Bioprocess and Biosystems Engineering manuscript No.
(will be inserted by the editor)

A dual parameter identification approach for data-driven predictive modelling of hybrid gene regulatory network-growth kinetics in *Pseudomonas putida* mt-2

Argyro Tsipa[1,†] · Jake Alan Pitt[2,3,†] · Julio R. Banga[2] · Athanasios Mantalaris[4]

Received: date / Accepted: date

Abstract Data integration to model-based description of biological systems incorporating gene dynamics improves the performance of microbial systems. Bioprocess performance, typically predicted using empirical Monod-type models, is essential for a sustainable bioeconomy. To replace empirical models, we updated a hybrid gene regulatory network-growth kinetic model, predicting aromatic pollutants degradation and biomass growth in *Pseudomonas putida* mt-2. We modelled a complex biological system including extensive information to understand the role of the regulatory elements in toluene biodegradation and biomass growth. The updated model exhibited extra complications such as the existence of oscillations and discontinuities. As parameter estimation of complex biological models remains a key challenge, we used the updated model to present a dual parameter identification approach (the "dual approach") combining two independent methodologies. Approach I handled the complexity by incorporation of demonstrated biological knowledge in the model-development process and combination of global sensitivity analysis and optimisation. Approach II complemented Approach I handling multimodality, ill-conditioning and overfitting through regularisation estimation, global optimisation and identifiability analysis. To systematically quantify the biological system, we used a vast amount of high-quality time-course data. The dual approach resulted in an accurately calibrated kinetic model (NRMSE:0.17055) efficiently handling the additional model complexity. We

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement 675585 (Marie Skłodowska-Curie ITN "SymBioSys"). JRB also acknowledges funding from the Spanish Ministry of Science, Innovation and Universities and the European Union FEDER under project grant SYNBIOCONTROL (DPI2017-82896-C2-2-R). Author JAP has been a Marie Skłodowska-Curie ESR at IIM-CSIC (Spain).
JAP is now at College of Engineering, Mathematics & Physical Sciences, University of Exeter, Devon, UK

E-mail: Prof A. Mantalaris at sakis.mantalaris@gatech.edu, or Prof J. R. Banga at julio@iim.csic.es
[†] These authors contributed equally to this work.
[1] Dept. of Civil and Environmental Engineering, University of Cyprus, 75 Kallipoleos Street, 1678 Nicosia, Cyprus
[2] (Bio)Process Engineering Group, Spanish National Research Council, IIM-CSIC, C/Eduardo Cabello 6, 36208 Vigo, Spain.
[3] RWTH-Aachen University Hospital, Joint Research Centre for Computational Biomedicine (JRC-COMBINE), 52074, Aachen, Germany.
[4]Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

tested model validation using three independent experimental data sets, achieving greater predictive power (NRMSE:0.18776) than the individual approaches (NRMSE I:0.25322, II:0.25227) and increasing model robustness. These results demonstrated data-driven predictive modelling potentially leading to bioprocess' model-based control and optimisation.

Keywords Bioprocess development · Parameter estimation · Complex biological systems · Predictive modelling · *Pseudomonas putida*

1 Introduction

Mechanistic dynamic models (also known as kinetic models) of non-linear systems, such as biological systems, are playing an increasingly important role in biotechnology and bioprocessing [1, 9, 49, 58]. Integration of data and model-based description of biological systems incorporating gene dynamics allow us to both quantitatively and systematically understand, extract meaningful knowledge and ultimately improve the performance of microbial systems and cell factories [1].

Bioprocess performance, which is typically modelled through unstructured and empirical Monod-type growth kinetic models, and scalability are essential to ensure sustainable growth of the bioeconomy (i.e. industrial biotechnology, biomedicine and synthetic biology). Monod-type models are based on black box approximations ignoring molecular interactions and gene regulation in each micro-organism. Consequently, they inaccurately predict bioprocess kinetics under different conditions [29, 44]. Furthermore, although systems biology approaches (in combination with omics and computational tools) seek to describe cellular behaviour in detail, certain limitations, such as application of Monod-type models to predict bioprocess kinetics result in lack of predictability and prevent their further applicability in bioprocess development [10, 30, 37].

Pseudomonas putida is a metabolically versatile soil bacterium, with biotechnological capacities ranging from chemical production [3] to mineralisation of a large number of industrially important aromatic pollutants [39]. It is considered a cell factory platform for producing targeted chemicals through metabolic engineering [19] and a workhorse in synthetic biology [36]. The strain mt-2 contains the TOL plasmid, which represents a paradigm of global and specific gene regulation [41]. The TOL gene regulatory network (GRN) consists of the *Pr*, *Ps*, *Pu* and *Pm* promoters which control the biodegradation of toluene and *m*-xylene pollutants [52]. Modelling of gene dynamics of the TOL network facilitates model analysis towards gene regulatory network optimisation, model-based control and engineering strategies. Towards this direction, Koutinas et al. 2010 [28] modelled the regulatory logic of *Pr*/*Ps* node upon *m*-xylene oxidative catabolism, providing a systemic understanding of the causality and connectivity of the regulatory elements.

Koutinas et al. (2011) [27] extended their previous framework to capture *m*-xylene biodegradation and biomass growth through modelling of the TOL GRN. Specifically, the bioprocess performance was predicted through the GRN model which informed the formulation of growth kinetics, leading to *m*-xylene biodegradation and biomass growth prediction through gene dynamics. Toluene is the primary *P. putida* mt-2 substrate [52] and it was observed that the *m*-xylene-based Koutinas et al. (2011) [27] model inaccurately captures TOL promoters behaviour and bioprocess kinetics upon toluene biodegradation. Tsipa et al. (2016, 2017, 2018) proved that toluene efficiently activates the *ortho*-cleavage GRN [53–55]. This activation occurred due to enzyme activity encoded by the TOL pathway which led to toluene biotranformation, triggering expression of BenR protein [12] of the *ortho*-cleavage path-

way. BenR is controlled by the *ortho*-cleavage *PbenR* promoter [13] and it is a transcriptional regulator of both *ortho*-cleavage *PbenA* and TOL plasmid *Pm* promoters. Therefore, BenR expression interconnects the TOL and *ortho*-cleavage regulatory networks triggering the metabolic cascade of the latter [31]. This interconnectivity is the prerequisite step for Krebs cycle metabolites formation which are essential for biomass growth.

The development of kinetic models is a complex procedure with a number of different stages starting with model structure development and parameter estimation. Usually, this model-building process is implemented as an iterative closed-loop procedure [4, 25] whereby once a model structure has been proposed, parameter estimation (i.e. model calibration) must be performed, using the available experimental data. Parameter estimation in nonlinear dynamic models is a challenging inverse problem [51, 56]. In a frequentist framework, the problem is usually formulated as a maximum likelihood estimation, which takes the form of a nonlinear optimisation problem, subject to dynamic and algebraic constraints. This class of problems has many potential pitfalls due to frequent non-convexity (multimodality), ill-conditioning and lack of identifiability [11, 24, 32, 35, 56]. The typical ill-conditioning of these problems originates from (a) lack of information due to scarce and noisy data and (b) identifiability issues due to over-parameterisation of the models. Further, even when good fits are obtained, one should be aware of possible overfitting (i.e. fitting the noise rather than the signal).

In addition to the aforementioned challenges in parameter estimation problems of kinetic models two extra complications exhibited in the updated model are considered here: (a) the presence of oscillatory behaviour [53, 55] and (b) the existence of discontinuities (both explicit and implicit) in the model. Models that can fit oscillatory behaviour are generally remarkably flexible and thus rather prone to overfitting [40]. A clear example of this issue is the difficulty in finding a unique period, or frequency for the oscillations. In many cases, it is possible to represent the same data set equally well by simply changing the frequency of the oscillations. Discontinuities introduce non-smoothness into the optimisation function, creating additional difficulties for optimisation solvers.

In this paper, the hybrid GRN-growth kinetic model of Koutinas et al. (2011) [27] is updated to capture toluene biodegradation, including more biological information at the gene level and integrating a vast amount of data. This way a more comprehensive and universal understanding of the causality and connectivity of the regulatory elements coupled with bioprocess performance prediction is achieved. We also introduce a simple mathematical characterisation connecting gene regulation machinery to bacterial lag phase. Lag phase was not modelled by Koutinas et al. (2011) [27]. Development of biologically relevant lag phase model equations is challenging due to limited physiological knowledge. Therefore, most modelling definitions of lag phase are mathematical or geometric [5, 34].

We address the parameter estimation challenge using a dual parameter identification approach, the dual approach, which combines two methods: (I) the biologically inspired and (II) the purely statistical. Approach I follows a model-building process based on the existing biological knowledge, exploiting global sensitivity analysis and optimisation. Approach II meets Approach I when the model structure is finalised and is based on a novel regularisation estimation, exploiting global optimisation and identifiability analysis [40]. The dual approach results in prediction envelopes, defined by the individual estimates. To test the predictive power of the calibrated model and detect issues such as overfitting, we adopt a data-driven validation scheme. Specifically, the available data are partitioned into two sets. The first set is used to calibrate the model (i.e. the parameter estimation problem, also called training). Then, the predictive quality of the calibrated model is tested utilising the second

independent set in which we use three independent experimental data sets.

2 Methods

2.1 Problem statement

We consider dynamic models of biosystems, described by sets of nonlinear ordinary differential equations (ODEs), with discrete events in the state space. In particular, we study models of the following type:

$$\frac{d\mathbf{x}(t,\theta)}{dt} = \mathbf{f}_i\left(t,\mathbf{u}(t),\mathbf{x}\left(t,\theta\right),\theta\right) \; for t_{i-1} < t < t_i \tag{1}$$

$$\mathbf{y}\left(\mathbf{x},\theta\right) = \mathbf{g}\left(\mathbf{x}\left(\theta,t\right),t,\theta\right) \tag{2}$$

$$\mathbf{x}\left(t_0,\theta\right) = \mathbf{x}_0 \tag{3}$$

Where $\mathbf{x} \in \mathbb{R}^{N_\mathbf{x}}$ represents the states of the system as time-dependent variables under the initial conditions $\mathbf{x}_0$, $\theta \in \mathbb{R}^{N_\theta}$ is the parameter vector, $\mathbf{u}(t)$ represents the vector of time-dependent inputs (e.g. stimuli) affecting the system and $t \in [t_0, t_{end}] \subset \mathbb{R}$ is the time variable. Each t_i represents the switching time from the ith explicit discontinuity, with each f_i representing the nonlinear function during that specific period of time. The observation function $g : \mathbb{R}^{N_\mathbf{x}\times N_\theta} \mapsto \mathbb{R}^{N_\mathbf{y}}$ maps the states to a vector of observables $\mathbf{y} \in \mathbb{R}^{N_\mathbf{y}}$, that can be measured. Each implicit discontinuity also has a switching condition given by:

$$\mathbf{h}_j\left(\mathbf{x}\left(t_j,\theta\right)\right) = 0 \tag{4}$$

Where the jth implicit discontinuity is switched at the time t_j when the condition is satisfied. Here for the jth implicit discontinuity, we have a change function given by:

$$\mathbf{v}_j\left(t_j,\mathbf{x}\left(t_j,\theta\right),\theta\right) = \mathbf{x}\left(t_j^+,\theta\right) - \mathbf{x}\left(t_j^-,\theta\right) \tag{5}$$

Where $t_j^+ = \lim_{\varepsilon\to 0} t_j + \varepsilon$ and $t_j^- = \lim_{\varepsilon\to 0} t_j - \varepsilon$. We now consider the parameter estimation problem for the aforementioned discontinuous nonlinear dynamic system. We formulate this problem as an optimisation problem, where we seek to optimise the maximum likelihood cost function given by:

$$L\left(\tilde{\mathbf{y}}|\theta\right) = \prod_{k=1}^{N_e}\prod_{j=1}^{N_{y,k}}\prod_{i=1}^{N_{t,k,j}} \frac{1}{\sqrt{2\pi\sigma_{kji}^2}} \exp\left(\frac{-\left(y_{kji}\left(\mathbf{x}\left(t_i,\theta\right),\theta\right) - \tilde{y}_{kji}\right)^2}{2\sigma_{kji}^2}\right) \tag{6}$$

Where N_e is the number of experiments, $N_{y,k}$ the number of observables within the said experiment and $N_{t,k,j}$ is the number of time points within each observable. Additionally, $\tilde{y}_{kji}$ represents the measured behaviour of the ith time point, of the jth observable, in the kth experiment and with σ_{kji} representing the standard deviation of said point.

(7)

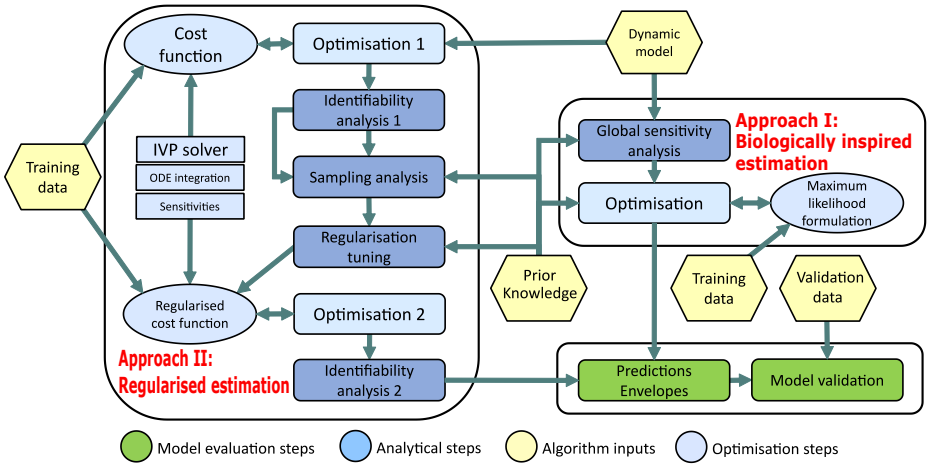


Fig. 1 An overview of the dual parameter identification procedure followed.

2.2 Model Development

The integrated work-flow for the combined method is shown in Fig 1. We use Approach I to re-develop the Koutinas et al. (2011) [27] model using the biological knowledge obtained from the literature and the systematic evaluation of the *Pr*, *Ps*, *Pu*, *Pm*, *PbenR* and *PbenA* transcriptional kinetics of the interconnected TOL and *ortho-* cleavage GRNs [53]. The model structure is based on Fig 2, which represents the interconnected TOL and *ortho-* cleavage regulatory networks as a set of genetic circuit logic gates signals transmitted between different molecular components [57]. These specified combinations of the logic gates facilitate simpler descriptions of the current inherent regulatory loops. Hill functions were employed as input functions to the genes, enabling dynamic characterisation of bioprocess components [2]. The model development process based on biological knowledge is explained below and the mathematical representation of the model is described in Table 1.

Upon toluene entry, the inactive form of dimer XylR ($XylR_i$) is oligomerised to the hexamer $XylR_a$ and becomes transcriptionally competent [41]. Both inactive and active XylR forms act as auto-repressors of XylR expression, down-regulating *Pr* expression [6] (Equation 8). The *Pr* expression however, restores to the basal level for toluene concentrations below 0.3 mM, possibly because XylR stops down-regulating *Pr* [53]. Therefore, we diverge from Koutinas et al. 2011 [27] for toluene concentrations below the 0.3 mM threshold (Equation 9). $XylR_a$ dynamically equilibrates with $XylR_i$ [15] (Equation 10 [27]).
Tsipa et al. (2016) [53] observed that *Ps* and *Pu* mRNA expression peaked at 60 mins, suggesting completion of equilibrium between $XylR_i$ and $XylR_a$ below this time point and $XylR_a$ subsequent degradation (Equation 11). We modify the $XylR_a$ association rate constant of Koutinas et al. 2011 [27] to depend on initial toluene concentration (Equation 12) as shown by [53]. *Ps* expression is modelled by *Ps1* promoter using Equations 13, 14 [7, 27]. XylS dimerises to become transcriptionally active and up-regulates *Pm*; we model the dynamic equilibria between inactive ($XylS_i$) and active ($XylS_a$) forms diverging from Koutinas et al. 2011 [27] as the *Pm* depends on the initial toluene concentration [53]. We express the association rate constant as shown in Equation 15. Equations 16, 17 express the inactive and active forms of XylS [27].
Pu is expressed in Equations 18, 19 [27]. Although all the enzymes encoded in the *upper*

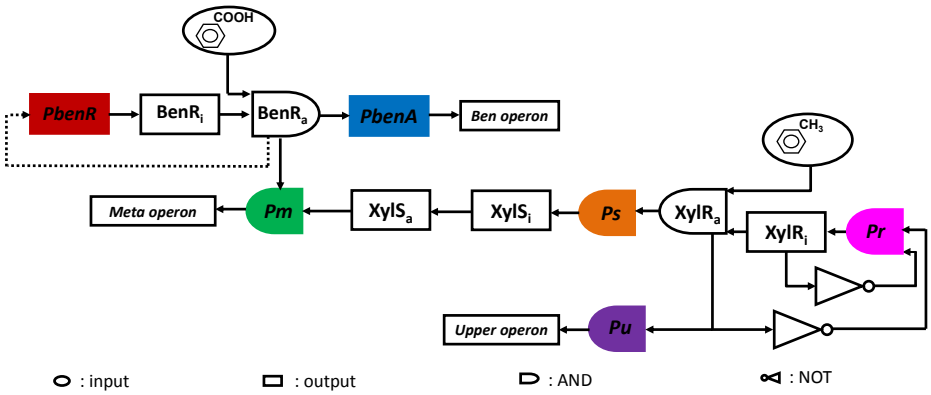


Fig. 2 Upon toluene entry, the inactive form of XylR ($XylR_i$) oligomerises forming the active molecule $XylR_a$ which activates *Pu* and *Ps* promoters. Both *XylR* forms down-regulate their own promoter, *Pr*. Upon *Pu* activation the genes of the *upper* operon encode for the enzymes which catalyse toluene catabolism. *Ps* activation and toluene catabolism lead to overexpression of the *xylS* gene dimerising the inactive *XylS* protein ($XylS_i$) to the active protein form ($XylS_a$). *XylS* dimerisation activates the *Pm* promoter. In the chromosomal pathway, *PbenR* controls *benR* gene transcription, which encodes for the inactive *BenR* protein form ($BenR_i$). Toluene catabolsim activates *BenR* ($BenR_a$) which up-regulates the *Pm* promoter of *TOL* and *PbenA* of the chromosome. *PbenA* controls *ben* operon transcription which encodes for the enzymes responsible for further transformation of toluene to Krebs cycle metabolites necessary for biomass growth.

operon are expressed, one has been considered as the rate-limiting and dominates on controlling the pathway, XylU [27] (Equation 20). Toluene biotransformation activates the BenR protein of the *ortho*-cleavage pathway [12]. BenR, like XylS belongs to AraC family [16], so we assume that BenR dimerises, becomes transcriptionally active ($BenR_a$), and exists in dynamic equilibrium with its inactive form ($BenR_i$) (Equation 21). We express the association rate constant between $BenR_i$ and $BenR_a$ as shown in Equation 22, due to dependence on toluene initial concentration [53].

PbenR and *Pm* expression peaks at 90 min [53]. *Ps* and *Pu* expression, which are both activated by $XylR_a$, reached their maximum at 60 min [53]. Under different growth conditions, Marques et al. (1994) [33] found that both *Ps* and *Pu* reach maximal expression simultaneously after 10 mins of induction. We, therefore, assume that *Pm* and *PbenR* are both triggered by $BenR_a$, suggesting auto-up-regulation of BenR protein (Equation 23). Equation 24 proposes a dynamic equilibrium between $BenR_i$ and $BenR_a$ in the first 90 mins, followed by $BenR_a$ dissociation after 90 mins.

$BenR_a$ up-regulates *Pm* [12]. So, we modify the Koutinas et al. (2011) [27] *Pm* function to capture both $BenR_a$ and $XylS_a$ regulatory mechanism (Equations 25, 26). *Meta*-operon encodes for enzymes catabolising further toluene into Krebs cycle metabolites through the TOL pathway. It has been assumed that an enzyme expressed by the *meta*-operon is rate-limiting, naming it XylM [27] (Equation 27).

BenR up-regulates *PbenA* which is modelled as seen in Equations 28, 29. *PbenA* expression controls *ben* operon expression (*ben*ABCDKX) [48] assisting on further catabolism to Krebs cycle intermediates. Therefore, we assume that a *ben* operon enzyme is the rate-limiting of the *ortho*-cleavage pathway, calling it BenB (Equation 30).

Microbial growth kinetics was linked to the updated GRN by focusing on the enzymatic steps of the GRN model that catalyse substrate degradation and biomass growth. Toluene biodegradation kinetics is dependent on XylU enzyme (Table 2, equation 31, 32). Instigated

by biomass growth co-dependence on the expression of the XylM and BenB rate-limiting enzymes, we reformulate the mathematical representation of the specific biomass growth rate (Table 2, equation 33,34) taking into account the decay rate (Table 2, Equation 35). The parameters of the model are presented as supplementary material (Tables S1.2).

2.3 Oscillatory *Pm* behaviour

The oscillatory behaviour of *Pm* has been observed at population level at toluene concentration threshold above 0.9 mM [53, 55]. There is currently no existing biological explanation for specific *Pm* oscillatory expression in the literature. The most common cause of oscillations is a negative feedback loop [2]. Here, we propose and explore the consequences of a scenario where *Pm* participates in a negative feedback loop with 2 transcription factors (R and I) of another pathway(s) (Fig 3). In a scenario based on Ferrell et al. 2011 [20] the transcription factor R down-regulates *Pm* expression and I is between *Pm* and R. We hypothesise that *Pm* controls I gene expression leading to I expression, which in turn up-regulates R gene expression encoding for R protein whose expression down-regulates *Pm* (Fig 3). The ODE Eqs (Table 2, 36 – 39) describe the promoter of the oscillatory behaviour, as sustained limit cycle oscillations. We assume mass action kinetics combined with Hill functions for the response between transcription factors and *Pm* and depolymerisation activation of I and R.

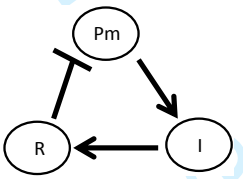


Fig. 3 The proposed scenario, describing the oscillatory behaviour of *Pm* via a negative feedback loop. I: intermediate protein, R: regulatory protein of *Pm*.

2.4 Mathematical modelling of lag phase

We propose a simple lag phase description based on maximum expression of the promoters controlling toluene bioconversion (Eq 40). Where l_{TOL} is the time point (60 minutes) where the *Pu* promoter of TOL peaked [53], l_{chrom} is the time point (90 minutes) where the *PbenR* promoter of the *ortho-cleavage* pathway of the chromosome (which interacts with TOL) peaked [53] and tol_{INI} is the initial toluene concentration in the total count. Eq 40 indicates that we can estimate lag phase duration (at least for this system), using the time point where key promoters are maximally expressed. This demonstrates the critical role of the promoters' kinetics in the bioprocess.

2.5 Experimental methods

The experimental data were generated using the following methods (which are described in detail in [53]): (i) microbial cultivation, (i) substrate and biomass analyses, (iii) prepara-

tion and isolation of total RNA, cDNA Synthesis and qPCR. The qPCR results were statistically analysed. Three independent cultures were grown at each condition tested, while the promoters of the mRNA expression were measured in triplicates for each time point. For each promoter, the average mRNA expression and its standard deviation were calculated. The experimental error bars for promoter expression at each time point were derived from three independent (biological) replicates and three qPCR internal (technical) replicate measurements. One way ANOVA (SigmaStat version 3.5, Systat Software UK Ltd, UK) was conducted, to clarify significant differences in the mRNA expression profiles of all the promoters. P-values were calculated through comparison of the average mRNA expression between two given time points. The level of significance was accepted at P-values lower than 0.05.

3 Parameter identification with a dual approach

3.1 Approach I: Biologically inspired parameter estimation

3.1.1 Solution strategy

Herein, we present a solution strategy which is based on the framework described in detail by Kiparissides et al. (2011) [25]. Collecting experimental data *in-vivo* is combined with *in-silico* approaches to model biological systems resulting in a closed loop approach between *in-silico* and *in-vivo*. In this closed loop, developing a mathematical model describing biological phenomena is followed by global sensitivity analysis and parameter estimation. The model is built and tested using one set of data. Another set of data is used for global sensitivity analysis and parameter estimation. Once the model is complete, its predictive capability is tested with independent experimental data sets. This leads to a mathematical bioprocess representation, with an adequate predictive capability and model validation. [25,26]. The model simulation and parameter estimation processes were implemented in the process modelling environment gPROMS® (Process Systems Enterprise, 2014) and were computed on an Intel Core i7-2600 PC with 8GB RAM running Windows 7.

3.1.2 Global sensitivity analysis

Global Sensitivity Analysis (GSA) was performed in order to systematically determine output variability (due to the nonlinearities in the model) and evaluate the relative influence of the uncertainty of the parameters in the outputs. The observed outputs of the model are: *Pr*, *Ps*, *Pu*, *Pm*, *PbenR*, *PbenA*, substrate (*tol*) and biomass (*X*). Nominal values from the Koutinas, et al. 2011 [27] model initialised the GSA. For the nominal BenR synthesis, the parameter values of the *ortho*-cleavage regulatory network in Koutinas, et al. 2011 [27] of XylS synthesis were used because of the functional similarities [12]. The nominal *PbenR* and *PbenA* expression was set to nominal *Pm* values, since *PbenR* and *PbenA* are regulated by BenR analogously to *Pm* by XylS. The nominal value of transition parameter values k_{xylRa} and k_{BenRa} was set to 10. Sobol's method [50] was used for GSA; the method was implemented in Matlab and connected to gPROMS via goMATLAB. Parameter importance was established using sensitivity indices (SI) ranging from 0 (low significance) to 1 (high

significance); it is assumed that parameters with SIs higher than 0.1 (which corresponds to 10% experimental error) are significant [47] . The sensitivity indexes were calculated on the GUI-HDMR [59] Matlab package. GSA identifies the significant model parameters, initialising the parameter estimation process. The random intervals used were 50000 and the nominal values were ranged ± 20. The time intervals examined were selected before and after the 60 and 90 minutes where the different behaviour of the promoters *Ps*, *Pu* and *PbenR* was observed and at later time points. Specifically, these time points were 50, 70, 100 and 400 minutes.

3.1.3 Optimisation

The maximum likelihood formulation (Eq 6) can be reformulated to the maximal log function (Eq 8), when we assume independent normally distributed errors with zero means and constant standard deviation. The parameters were estimated by optimising the maximal log function. This, simultaneously provides parameter estimation in both the physical process model and the measuring instrument variance model. The objective function maximises the probability that the model predicts experimental measurements [28].

$$\frac{n}{2}\ln(2\pi)+\frac{1}{2}\min_{\theta}\left\{\sum_{k=1}^{N_e}\sum_{j=1}^{N_{y,k}}\sum_{i=1}^{N_{t,k,j}}\left(ln\left(\sigma_{kji}^2\right)+\left(\frac{y_{kji}\left(\mathbf{x}\left(t_i,\theta\right),\theta\right)-\tilde{y}_{kji}}{\sigma_{kji}}\right)^2\right)\right\} \tag{8}$$

$$\sigma^2=\omega^2 \tag{9}$$

Subject to the system described in Eqs 1 - 5 and the parameter bounds described in 12, with n being the total number of data points. We denote the solution of this optimisation as ϑ. The above formulation can be reduced to a recursive least squares parameter estimation, if no variance model for the sensor is selected. The constant variance used in the optimisation process is calculated by Eq 9. The constant variance of the experimental results at each time point was set to 0.1. Following GSA, the significant parameters are estimated in gPROMS.

3.2 Approach II: Regularised estimation

3.2.1 Solution strategy

Here, we use a solution strategy designed to deal with both the issues of multimodality and overfitting [40]. We use a hybrid optimisation solver, combining the robustness of global optimisation with the efficiency of local optimisation and regularisation methods.

The resulting procedure is based on a two-step optimisation approach: (i) an initial exploratory optimisation is performed and subsequently used to tune the regularisation term, (ii) a second optimisation stage is used to find the regularised global minimum. The first optimisation step is used as an efficient sampling procedure for tuning the regularisation. Then, the second optimisation step is used to find the actual solution of the estimation problem. The initial optimisation step uses no prior knowledge and all the sampled points in the parameter space are saved and subsequently analysed. The resulting information is used to tune the regularisation term in the second optimisation step.

Parameter values that are not identifiable can take almost any value inside a large feasible range without having an impact to the cost function value (i.e. the model fit is independent

of their value). In order to avoid this effect, we perform a practical identifiability analysis after the exploratory optimisation, to judge which parameter samples need to be combined with prior information (e.g. available in the literature, or from previous knowledge of the experimentalists). Parameters that are deemed to be unidentifiable are assigned a reference value including this prior information.

3.2.2 Optimisation 1

Under specific conditions [45], the maximisation of the likelihood formulation is equivalent to the minimisation of the weighted least squares cost given by:

$$Q_{NLS}(\theta) = \sum_{k=1}^{N_e} \sum_{j=1}^{N_{y,k}} \sum_{i=1}^{N_{t,k,j}} \left(\frac{y_{kji}(\mathbf{x}(t_i,\theta),\theta) - \tilde{y}_{kji}}{\sigma_{kji}} \right)^2 = \mathbf{r}(\theta)^T \mathbf{r}(\theta) \tag{10}$$

The parameter estimation problem can, therefore, be posed as the following minimisation of the weighted least squares cost:

$$\min_{\theta} Q_{NLS}(\theta) = \mathbf{r}(\theta)^T \mathbf{r}(\theta) \tag{11}$$

Subject to the system described in Eqs 1 - 5, with the additional parameter bounds:

$$\theta_i^{min} \leq \theta_i \leq \theta_i^{max} \, \forall \, \theta_i \in \theta \tag{12}$$

The optimal parameter vector $\hat{\theta}$ that solves the above problem is the maximum likelihood estimate of the model parameters.

In the first optimisation step, we minimise the following cost function:

$$\begin{aligned} \widehat{Q}_{NLS}(\theta) &= \sum_{k=1}^{N_e} \sum_{j=1}^{N_{y,k}} \sum_{i=1}^{N_{t,k,j}} \left(\frac{N_{kj}\left(y_{kji}(\mathbf{x}(t_i,\theta),\theta) - \tilde{y}_{kji}\right)}{\sigma_{kji}} \right)^2 \\ &= \widehat{\mathbf{r}}(\theta)^T \widehat{\mathbf{r}}(\theta) \end{aligned} \tag{13}$$

The above (Eq 13) is the log likelihood cost, i.e. the difference between predictions and experimental data, weighted by their individual standard deviations. $\widehat{Q}_{NLS}$ is a version of Q_{NLS} (Eq 10), where we normalise the residuals for each observable, using the multiplier $N_{kj} = \frac{\max(\sigma_{kj})}{\max(\tilde{\mathbf{y}}_{kj})}$ to avoid a bias due to scaling.

$$\min_{\theta} \widehat{Q}_{NLS}(\theta) = \widehat{\mathbf{r}}(\theta)^T \widehat{\mathbf{r}}(\theta) \tag{14}$$

This minimisation is subject to Eq 1 - 5, with the additional parameter bounds:

$$\theta_i^{min} \leq \theta_i \leq \theta_i^{max} \, \forall \, \theta_i \in \theta \tag{15}$$

Where we denote the solution to the optimisation as $\widehat{\theta}$. For this minimisation subject to the dynamic constraints, we adopt a single-shooting procedure, where we need to solve the initial value problem (IVP, i.e. the dynamics and initial conditions) for each evaluation of the objective function in the outer optimisation. The IVP is solved using *AMICI* [21],

a high level wrapper for the *CVODES* solver [46], currently regarded as state of the art. One major advantage of *AMICI* is its capability to efficiently deal with discontinuities. For the outer minimisation we use the *enhanced scatter search* (*eSS*) [18] approach. *eSS* is a hybrid meta-heuristic, combining aspects of both global (non-deterministic) and local (deterministic) optimisation solvers. As the local optimiser within *eSS* we use *NL2SOL* [14], an efficient gradient based solver. The results presented here correspond to the *MEIGO* [17] toolbox implementation of eSS. We provide *NL2SOL* with accurate (local) sensitivity information, calculated via *AMICI*, which results in faster and more robust convergence, in comparison to a finite difference calculation. The gradient based nature of *NL2SOL* allows us to exploit *AMICI*'s strength dealing with discontinues.

In this initial optimisation step, we assume no prior knowledge of the system (with the exception of parameter bounds) and focus on the problem from a purely computational standpoint. During this process, we save all the evaluated parameter points together with their cost function value, and we re-use this information in the posterior sampling analysis procedure. We save every parameter point $\theta_{S,i}$ selected by *eSS* during our exploratory optimisation stage, along with the cost $\widehat{Q}_{NLS}(\theta_{S,i})$ associated with the parameter point. We create a matrix Θ (Eq 16), where each column is a parameter vector (selected in the optimisation) and a vector ζ (Eq 17) with the cost for each of these parameter vectors.

$$\Theta = \left[\theta_{S,1}, \ldots, \theta_{S,N_S}\right] \in \mathbb{R}^{N_\theta} \times \mathbb{R}^{N_S} \tag{16}$$

$$\zeta = \left[\widehat{Q}_{NLS}(\theta_{S,1}), \ldots, \widehat{Q}_{NLS}(\theta_{S,N_S})\right] \in \mathbb{R}^{N_S} \tag{17}$$

Where N_S is the number of parameter points selected by *eSS*.

3.2.3 Sampling analysis

We incorporate this latter type of prior knowledge (i.e parameter values and/or certain aspects of the dynamic behaviour of the system) by analysing the samples obtained during our initial parameter estimation. In particular, we scan the sample for points in parameter space that are in agreement with the prior knowledge about the dynamics (e.g. the frequency of the oscillations of certain states). We then select the point that has the lowest cost from this subset of parameter points as a reference achieved via sampling. We denote the set of all possible dynamics that fit our prior knowledge with respect to the dynamics as $\mathbf{X}_D$. We then reduce our parameter sample set Θ to the subset Θ_D (Eq 18) that coincides with $\mathbf{X}_D$. We also reduce our cost vector in the same way (Eq 19).

$$\Theta_D = \{[\theta_{S,1}, \ldots, \theta_{S,N_D}] \subset \Theta : \mathbf{x}(t, \theta_{S,i}) \in \mathbf{X}_D \, \forall \, \theta_{S,i} \in \Theta_D\} \tag{18}$$

$$\zeta_D = \{\widehat{Q}_{NLS}(\theta_{S,i}) \in \zeta \, \forall \, \theta_{S,i} \in \Theta_D\} \in \mathbb{R}^{N_D} \tag{19}$$

Where N_D is the number of parameter points within the sample that match the dynamics classified in $\mathbf{X}_D$. We now find the parameter value within our reduced sample with the lowest cost, i.e. the parameter vector with the lowest cost that matches the expected dynamics. This is used as the sample input into the reference parameter (Eq 23).

$$\theta_S^{ref} = \theta_{S,i} \in \Theta_D : \zeta_{D,i} = \min(\zeta_D) \tag{20}$$

3.2.4 Identifiability analysis 1

During our initial parameter estimation, we have disregarded all prior knowledge about the system, likely leading to said estimation resulting in an overfit solution. Therefore, we will use a regularisation term in the second minimisation, for which we need a reference parameter vector (i.e. a good guess). For the particular bioprocess considered here, some prior knowledge for parameter values can be found in the literature [27] but not enough to assign a reference value to every parameter. We, therefore, use our previously obtained sample of parameter space to generate the rest of the information needed. In this step, we also take into account identifiability issues. Due to a lack of identifiability, some parameters can take any value without changing the corresponding cost function, this will be reflected in the sample obtained in the first step. It should be noted that the size and complexity (especially the discontinuities) of our model restricts the effectiveness of structural identifiability analysis tools. Thus, we use the*VisId* [22] toolbox to perform a practical identifiability analysis. *VisId* computes collinearity indices between sets of parameters to determine, which subsets can be uniquely identified and uses sensitivity information around one particular parameter point to perform its analysis.

$$V(\theta_V) : \frac{\partial \mathbf{x}(t, \theta_V)}{\partial \theta} \mapsto \theta_{Iden} \, for \, a \, given \, \theta_V \tag{21}$$

VisId maps (denoted here as V) the sensitivities (of the states with respect to the parameters) calculated at a given point in parameter space to the set of identifiable parameters, as described in [22]. Here (Eq 22), we use our first (overfit) solution $\widehat{\theta} \in \Theta$ as this parameter vector and calculate the sensitivities at said point using *AMICI*. As a result of this identifiability analysis, we determine which parameters are identifiable and use this information, together with the sampling, to find the reference vector to be used in the regularisation term. From our identifiability analysis we find a set $\widehat{\theta}_{Iden}$ containing the parameters that are deemed identifiable by *VisId* at the point $\widehat{\theta}$.

$$V\left(\widehat{\theta}\right) : \frac{\partial \mathbf{x}\left(t, \widehat{\theta}\right)}{\partial \theta} \mapsto \widehat{\theta}_{Iden} \tag{22}$$

3.2.5 Regularisation tuning

We use our sampling reference parameter (Eq 20) as our reference values for all the parameters, where no information was available in the literature. When it comes to the parameters where we have found prior information in both the literature and our sampling procedure, we apply a slightly different approach. If a parameter has this prior information available and has been deemed identifiable in the previous step, we use the value from our sample. For the unidentifiable parameters, we combine (by averaging) the sample value with the literature value. In this way we balance the effect of lack of identifiability using the sampling approach. We take the prior knowledge information given in the literature and form a

vector θ_L^{ref}, we, then, form our reference parameter in the following way:

$$\theta_i^{ref} = \begin{cases} \frac{\theta_{S,i}^{ref} + \theta_{L,i}^{ref}}{2} & if\ \exists\, \theta_{L,i}^{ref} \in \theta_L^{ref}\ \&\ \theta_i \notin \widehat{\theta}_{Iden} \\ \theta_{S,i}^{ref} & otherwise \end{cases} \tag{23}$$

$$\forall\, \theta_i^{ref} \in \theta^{ref}$$

3.2.6 Optimisation 2

We perform a second minimisation using a Tikhonov regularisation term (Eq 24) with a normalisation factor (Eq 25) with respect to θ^{ref}, to avoid bias due to scaling. This second minimisation will produce a refined fit, without overfitting (i.e. a more generalisable one).We set up the second optimisation problem (Eqs 24 - 26) in a similar fashion to our original run (i.e. with log likelihood, *AMICI*, *eSS* and *NL2SOL*), the only exception being the addition of a regularisation term (Eq 24) in our cost function.

$$\Gamma(\theta) = (\theta - \theta^{ref})^T \mathbf{W}^T \mathbf{W} \left(\theta - \theta^{ref}\right) \tag{24}$$

$$\mathbf{W} = \begin{bmatrix} \frac{1}{\theta_1^{ref}} & 0 & \cdots & \cdots & 0 \\ 0 & \frac{1}{\theta_2^{ref}} & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{\theta_{N_\theta}} \end{bmatrix} \tag{25}$$

$$\min_{\theta} Q_R(\theta) = \widehat{Q}_{NLS}(\theta) + \alpha\Gamma(\theta) \tag{26}$$

Subject to the system described in Eqs 1 - 5 and the parameter bounds described in Eq 12. Where α is the regularisation parameter weighting the cost between the quality of fit and the effect of the regularisation, and the θ^{ref} parameter vector is the reference vector of parameter values obtained in the previous steps. We denote the solution to this optimisation as $\widetilde{\theta}$.

3.2.7 Identifiability analysis 2

Once we have achieved a regularised fit, we need to perform several analyses on the resulting calibrated model. In particular, we need to perform a new identifiability analysis using *VisId* at the new optimal (regularised) parameter values. We map our final optimal parameter vector $\widetilde{\theta}$ to the set of identifiable parameters at this point given by $\widetilde{\Theta}_{Iden}$ (Eq 27).

$$V\left(\widetilde{\theta}\right) : \frac{\partial \mathbf{x}\left(t, \widetilde{\theta}\right)}{\partial \theta} \mapsto \widetilde{\Theta}_{Iden} \tag{27}$$

4 Model evaluation

4.1 Envelope intervals

Due to the independent nature of the two different methods used, the two methods have independent bias. We look to exploit this by combining both methods into the dual approach as follows:

$$\widehat{\mathbf{y}}(\mathbf{x},\theta) = \frac{\mathbf{y}\left(\mathbf{x},\widetilde{\theta}\right) + \mathbf{y}(\mathbf{x},\vartheta)}{2} \tag{28}$$

Where $\widehat{\mathbf{y}}$ gives the behaviour observed by the dual approach, given by the midpoint of the predictions of the biologically driven and the regularised estimation calibration. From our dual approach we form not only a prediction but also a prediction envelope calculated by:

$$I_j = \left[min\left(\mathbf{y}\left(\mathbf{x},\widetilde{\theta}\right)\Big|_{t_j}, \mathbf{y}(\mathbf{x},\vartheta)\Big|_{t_j} \right),\ max\left(\mathbf{y}\left(\mathbf{x},\widetilde{\theta}\right)\Big|_{t_j}, \mathbf{y}(\mathbf{x},\vartheta)\Big|_{t_j} \right) \right] \tag{29}$$

Where I_j is the prediction interval for each $t_j \in [t_0, t_{end}]$ giving a smooth envelope over the time period. The prediction interval is the area between smallest and largest predictions given by the two methods used.

4.2 Metric of model calibration and validation

Using the normalised root mean square error (NRMSE, Eq 30) as a metric for quality of fit, we can assess our model on the quality of both calibrated fit and predictive power. For the assessment of predictive power the NRMSE is calculated based on all of the validation data sets available simultaneously. We use the NRMSE metric in order to have a general metric that is directly comparable, using the normalised version of the metric in order to avoid any bias due to scaling. This metric allows us to quantitatively measure the quality of both our fit and of our predictive power and can be used for direct comparison. We simulate our calibrated model under each of the conditions associated with each validation data set, in order both to calculate the residuals needed for the NRMSE metric and for a visual plotting (qualitative) of the model kinetics under each set of conditions.

$$NRMSE\left(\mathbf{y}\left(\mathbf{x}\left(t,\theta\right),\theta\right)\right) = \sqrt{\frac{\sum_{k=1}^{N_e}\sum_{j=1}^{N_{y,k}}\sum_{i=1}^{N_{t,k,j}}\left(\frac{y_{kji}-\tilde{y}_{kji}}{\max\left(\tilde{\mathbf{y}}_{kj}\right)-\min\left(\tilde{\mathbf{y}}_{kj}\right)}\right)^2}{n}} \tag{30}$$

Where n is the total number of data points.

5 Results

5.1 Identifiability Analysis and Global Sensitivity Analysis

The objective using the dual approach is not simply to produce a unique parameter vector, which explains the data (and performs well in additional validation tests). Instead, we seek a dual parameterer identification, which results in prediction envelopes (i.e. those defined by the two-state prediction trajectories of Approach I and Approach II). The dual approach combines the biological inspired (Approach I) and statistical (Approach II) parameter estimation methods, following well-specified procedures (as depicted in Fig 1). The Global sensitivity analysis (GSA) in Approach I and the identifiability analysis in Approach II provide significant information for each estimation and by extension to the dual fit and the prediction envelopes of the biological system studied.

Global sensitivity analysis (GSA) enables the connection of the mathematics with the system's biology. GSA identifies the significant parameters of the mathematical model, allowing us to initialise the parameter estimation (Supplementary material, S1.1.1, Figs S1.2, S1.2) and reduce model uncertainty (caused by a large number of model parameters). The GSA results reveal and demonstrate the biological relevance between parameters (inputs) and variables (outputs), confirming accurate model structure of such a complex biological system. For the biological system studied, GSA demonstrates that the significant parameters for *Pr*, *Ps*, *Pu*, *PbenR* and *PbenA* promoters are the ones associated with their expression. In addition, the parameters related to XylR and BenR synthesis are also found to be significant. BenR synthesis of the *ortho*-cleavage GRN is the main additional biological component that we added to the Koutinas et al. (2011) [27] model to increase biological information and model fidelity. The parameters related to BenR synthesis are found to be important for accurately expressing *Pm*, toluene and biomass concentration. This corroborates the importance of the *ortho*-cleavage pathway to model the transcriptional regulation and subsequently bioprocess kinetics. In the parameter estimation process, we prioritised estimation of the significant parameters as defined by GSA method.

Identifiability analysis explores a large number of parameters in complex biological models. The analysis is focused on the mathematical model without requiring any prior knowledge (i.e. parameter values, systems dynamics) of the modelled biological system and is performed as an analytical step in Approach II (S1.1.2, Figs. S1.3, S1.4). In the current model, some parameters are deemed to be unidentifiable due to collinearity between them and lack of sensitivity. It should be noted that partial lack of identifiability for a model of this complexity (21 states, but with only 8 observed, and more than 50 parameters) is not unexpected. Simply combining sets of independent estimations is not possible here as parameters form collinear groups within which relationships between parameters generate certain model outputs. Using the `VisId` analysis we find that there are 105 pairwise collinear relationships (Figs S1.5, S1.6), which means that there is a large lack of identifiability within the system. Simply combining parameter values can break these relationships, leading to artefacts. We circumvent this behaviour by instead combining the estimations of our two approaches directly.

5.2 Model calibration

Approach I is based on the model-building process which incorporates biological knowledge of the system while Approach II is based on the regularised estimation which when compared to the non-regularised estimation indicates more precise model calibration (S1.1.3).

Our approach results in a combined prediction trajectory, the dual fit, for each output (i.e. *Pr*, *Ps*, *Pu*, *Pm*, *PbenR*, *PbenA*, toluene and biomass), a prediction envelope to indicate the space of the predicted kinetics formed between the two individual approaches (Figure 4) and two sets of estimated parameter values (Table S1.2). The results demonstrate a good agreement between the calibrated model using the dual fit and the data (as shown by the NRMSEs values in Table 3). We note that the lag phase is also accurately predicted (Figure 4). The obtained envelope of the dual approach which captures *Pr*, *Ps*, *Pu* expression levels was narrow (Figure 4) suggesting an agreement between the individual approaches and the beneficiary effect of dual parameter identification to increase confidence over model structure process and purposes.

With respect to *Pm*, *PbenR* and *PbenA* simulated behaviour (Figure 4), the envelope space of the dual approach is broader as compared to the previous promoters expression. This could be due to a lack of identifiability of certain parameter, or to the increased variability when dealing with oscillations. Furthermore, *PbenR* and *PbenA* expression is not accurately represented in the model structure. *PbenR* has been modelled as being auto up-regulated as suggested by Tsipa et al. (2016, 2017) [53,55] but this transcriptional regulation mechanism has not been experimentally validated yet and may lead to insufficient model structure. Furthermore, *PbenA* is most probably co-up-regulated by both *ortho*-cleavage BenR and TOL XylS transcriptional factors [38, 53, 55]. The latter transcriptional regulation mechanism is not included in the developed model. The dual fit accounts for the shortcomings of each approach which results in an enhanced representation of the GRN system and provides more accurate representation of the mRNA expression patterns of both promoters (Fig 4) as compared to each individual approach. With respect to the oscillatory behaviour of *Pm*, despite the different trajectories predicted by each individual approach, the dual approach was able to represent the *Pm* calibration data well (Table S1.3). The purpose of the current modelling framework is to link the GRN to growth kinetics. Through this framework, the bioprocess is faced as an intracellular informatory process rather than a black box. We show that our dual approach, despite model deficiencies at the GRN level, was able to achieve a good fit to the experimental kinetic patterns of the bioprocess. The dual parameter identification accurately simulates toluene degradation and biomass formation calibration data (Table S1.3).

5.3 Model Validation

The best model fit to the calibration data is not always the best solution and can be misleading. We could have a case of overfitting, where the calibrated model would have poor predictive power for conditions different from the ones used in the calibration. To ensure that a model does not suffer from overfitting, its performance must be tested with a validation study (i.e independent data sets not used in the calibration). Herein, we have considered validation of a vast amount of independent data corresponding to experiments with three different initial conditions, namely initial toluene concentrations at 0.4, 0.6 and 1.2 mM which include systematic monitoring (i.e. every 30 min) of *Pr*, *Ps*, *Pu*, *Pm*, *PbenR* and *PbenA* promoters and toluene and biomass kinetics. These toluene concentration levels were chosen to represent extreme culture conditions, where the *P.putida* mt-2 culture may be unable to grow and thus the TOL and *ortho*-cleavage promoters may not be activated. Low levels (i.e. 0.4, 0.6 mM) could be an insufficient carbon source unable to support microorganism's growth, whereas high (i.e 1.2 mM) concentrations may be lethal due to the toxicity of toluene. The validation results are shown in Figs 5, 6, 7.

At every initial toluene concentration, the GRN model reproduces the magnitude and trend of all six promoters and recognises the different initial conditions. Furthermore, the dual

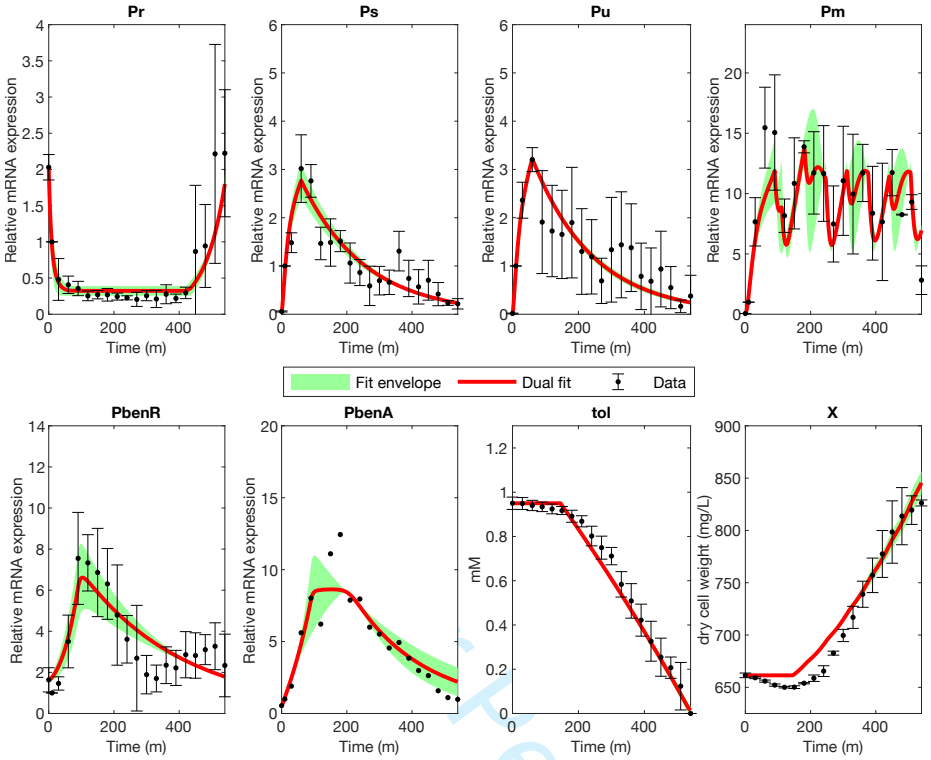


Fig. 4 A figure showing the simulation of the dual fit and envelopes resulting from the dual approach under the calibration conditions. The data here is the data used for the calibration of the model. The results for transcriptional kinetics are obtained as an average from nine individual measurements (three biological replicates and three technical replicates) at each point and the error bars are calculated for standard error. The results for substrates degradation and biomass formation are obtained as an average from three individual measurements at each point and the error bars are calculated for standard deviation

approach captures the *Pr*, *Ps* and *Pu* behaviour better than the individual approaches (NRMSEs shown in Tables S1.4-S1.6). For these variables, in every condition studied, the envelope space is narrow. Interestingly, the envelope space of the dual approach enclosing the trajectories of *PbenR* and *PbenA* is broader. These behaviours are under-estimated by the dual approach, indicating possible model deficiencies due to incomplete knowledge of the regulatory mechanisms involved. However, the dual approach results in a better representation and prediction of the experimental patterns, compared to the individual methods (NRMSEs shown in Tables S1.4-S1.6). With respect to the *Pm* predicted expression pattern, under the non-oscillatory conditions (Figs 5 - 6) , parameter α_{Pm}, which expresses the degradation of *Pm* mRNA, is responsible for the negative effect in the differential equation. The identifiability analysis performed with Approach II found this parameter to be practically unidentifiable, i.e. there is a lack of available information in the data to adequately estimate this parameter. This causes direct difficulty to capturing the state's decline. Approach I was able to better capture the decline. In the oscillatory case (i.e. 1.2 mM initial toluene concentration, Fig 7), the varying amplitude of experimental oscillations results in a discrepancy between the predicted fit of each individual approach. However, at every condition tested, the fit of the dual approach captures these dynamics more accurately compared to the indi-

vidual approaches (NMRSEs shown in Tables S1.4-S1.6). Overall, despite some prediction errors at the GRN level, the two most significant bioprocess design variables (i.e. toluene, and biomass) were adequately captured (NRMSEs shown in Tables S1.4-S1.6). The dual fit computed with the dual approach was able to capture the different patterns better than the individual approaches. The superiority of the dual approach in validation is quantitatively shown in Table 3, indicating a quality of validation similar to that of the calibration.

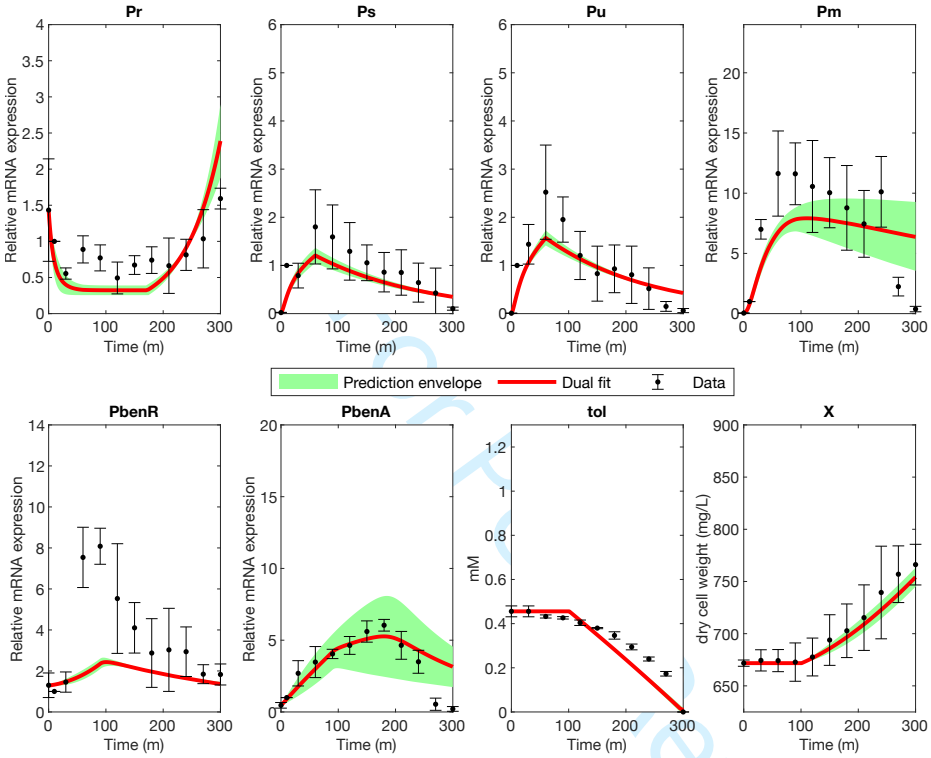


Fig. 5 A simulation of the dual fit and prediction envelopes under model validation conditions with initial toluene levels of 0.4 mM. The results for transcriptional kinetics are obtained as an average from nine individual measurements (three biological replicates and three technical replicates) at each point and the error bars are calculated for standard error. The results for substrates degradation and biomass formation are obtained as an average of three individual measurements at each point and the error bars are calculated for standard deviation.

6 Discussion

Parameter estimation of nonlinear kinetic models of biological systems is a crucial yet, currently, a rather intractable methodology. Traditional parameter estimation methods often result in an inaccurately calibrated model, mostly due to noisy data, over-parameterisation and strong nonlinearities (which can result in objective function with many local optima). Standardisation of this methodology could lead to robust models, assisting in optimal design, large-scale development, control and optimisation of bioprocesses.

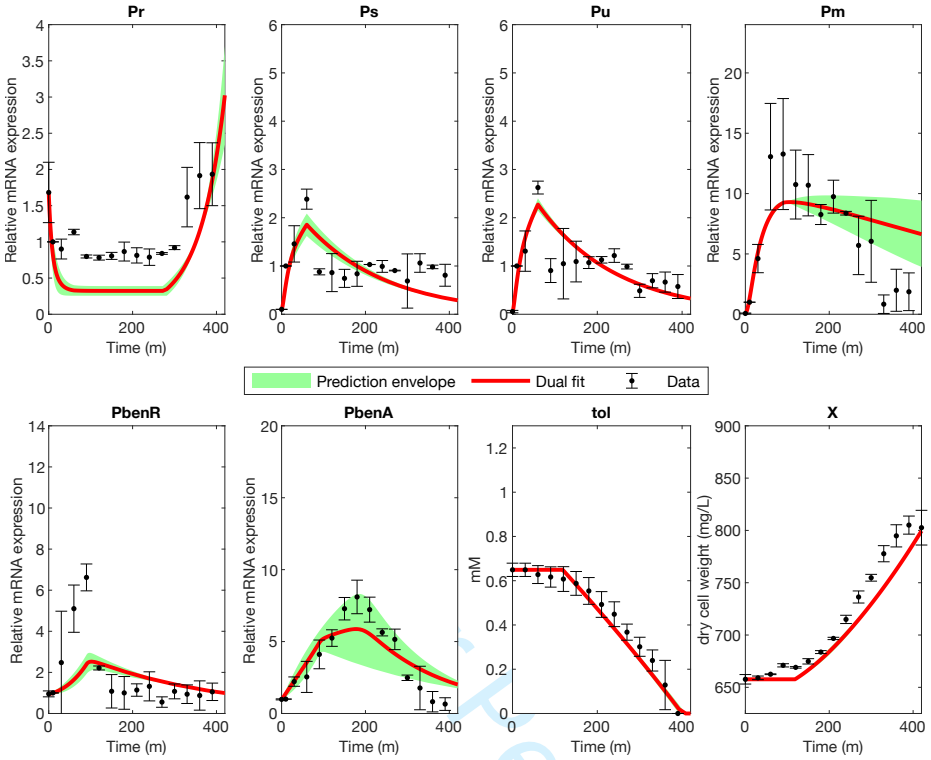


Fig. 6 A simulation of the dual fit and prediction envelopes under model validation conditions with initial toluene levels of 0.6 mM. The results for transcriptional kinetics are obtained as an average from nine individual measurements (three biological replicates and three technical replicates) at each point and the error bars are calculated for standard error. The results for substrates degradation and biomass formation are obtained as an average of three individual measurements at each point and the error bars are calculated for standard deviation.

We present the dual approach, a parameter identification method, combining mathematics with biology. The dual approach uses common mathematical tools and incorporates prior biological knowledge. Approach I allows the incorporation of demonstrated knowledge of the biological system in the model-development process handling any issues raised in the model structure due to the complexity of the biological system. Approach II is specifically intended to handle the issues of multimodality, ill-conditioning and overfitting complementing and verifying the parameter identification of Approach I.

We employ the dual approach in the hybrid GRN-growth kinetic model, which predicts bioprocess performance of aromatic pollutants degradation and biomass growth in *P.putida* mt-2 and is a paradigm of model and biological complexity. The dual approach exploits the advantages of each individual method, resulting in an accurately calibrated and validated kinetic model, which overcomes the challenges of parameter estimation of kinetic biological models. We improve the robustness of the model overcoming the many computational issues and pitfalls in modelling of complex biological systems. We are able to efficiently deal with the additional model complexity caused by multiple (implicit and explicit) discontinuities and oscillatory behaviour. Significantly, we increase the predictive capability of the model.

When dealing with prior information, one may be tempted to lean towards Bayesian ap-

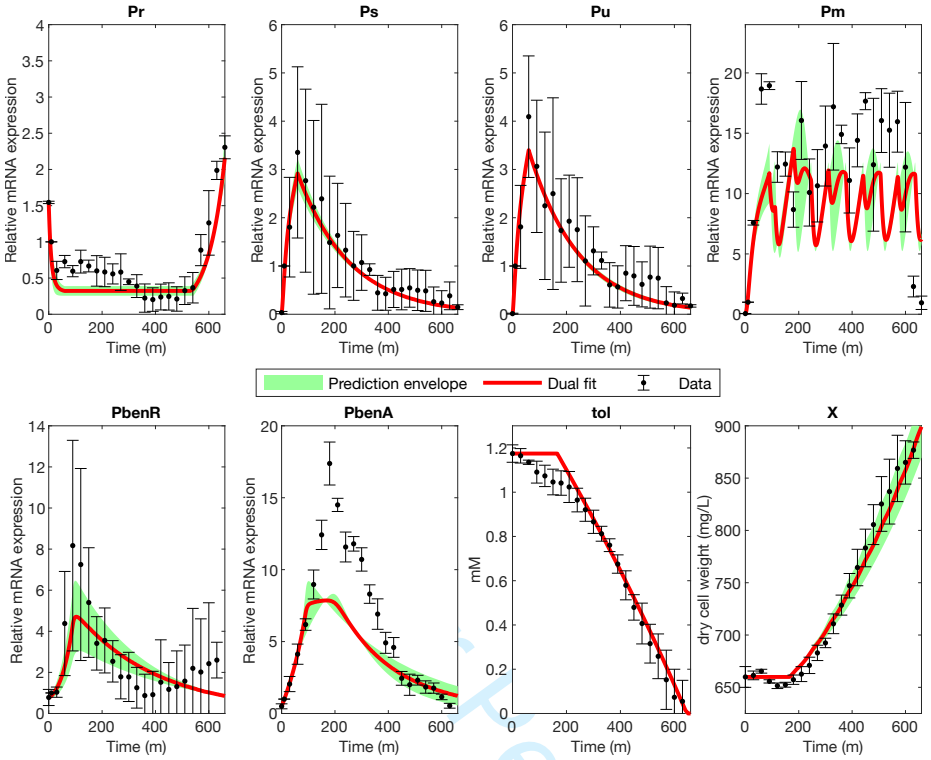


Fig. 7 A simulation of the dual fit and prediction envelopes under model validation conditions with initial toluene levels of 1.2 mM. The results for transcriptional kinetics are obtained as an average from nine individual measurements (three biological replicates and three technical replicates) at each point and the error bars are calculated for standard error. The results for substrates degradation and biomass formation are obtained as an average of three individual measurements at each point and the error bars are calculated for standard deviation.

proaches for parameter inference. However, here, we avoid using such methods due to the increased computational burden and known issues with the convergence of Bayesian algorithms in the presence of a lack of identifiability [42]. There is also a direct link between Bayesian estimation and regularisation [24], which we exploit in order to include our prior knowledge. One particular type of analysis that could be desirable is uncertainty quantification. Typically, this type of analysis is performed via the Fisher Information Matrix (or FIM) [23, 43], however, in the presence of a lack of identifiability, the FIM becomes ill-conditioned. This leads to numerical artefacts, and therefore the resulting uncertainty quantification becomes unreliable. A more rigorous method would be to use a bootstrapping approach. This, however, is highly computationally demanding [23] and extreme for the size and complexity of the model we have here.

Currently, the kinetic model-building process is lagging behind in producing high-throughput time-course reliable data [8]. Herein, we use a data-driven parameter estimation by means of high-quality time-series data through technical and biological triplicates, measured at 30-minute intervals, assisting in the efficiency of parameter estimation and model validation. In the dual approach, we combine practical identifiability analysis and GSA, which each have different purposes and complementary roles in experimental design and model-building

process. This may potentially lead to efficient utilisation and distribution of laboratory resources. Identifiability analysis (Approach II) is a useful mathematical tool to guide and enhance the experimental design, suggesting additional variables that should be measured to resolve any identifiability issues. In the current model, some parameters are unidentifiable suggesting that experimental data of additional states could be beneficial. Knowledge of the regulatory mechanisms of the complex biological system (Approach I) identifies the difficulty to measure such state variables, due to cost and time limitations of the existing experimental methods. Furthermore, unlike local sensitivity analysis, which is common in kinetic biological modelling [1], we employ GSA. The aim of GSA is to determine the degree of change of a model property such as gene expression in response to a change in the model parameters. As the parameters may represent quantities that can be manipulated by genetic engineering, such as protein concentrations, the analysis enables predictive links between potential targets and their effect on the behaviour of the biological system. For instance, in the model studied, based on the significance of the parameters of $XylR_a$ on Pr, Ps, Pu, and that of $BenR_a$ on Pm, a change on the structure of XylR and BenR protein may affect and accelerate the biodegradation process.

The complementarity of Approaches I and II enhances the purely mathematical background usefulness. One aspect of this is that Approach II explicitly uses the standard deviation of the experimental results in the estimation process. This complements the limitation of Approach I, which assumes independent, normally distributed measurement errors with zero means and constant variance. Further, model reduction could be applied to the studied model to decrease the complexity and therefore the computational burden of simulation. For instance, the nonlinearity caused by $XylR_a$ and $BenR_a$ could be described by a single equation, as the second equation describes a slower state of the biological system, which can be excluded as in the model described by Tsipa et al. (2018) [54]. Such decisions are usually dependent on a combination of factors such as (a) the purpose of the model and (b) the prior knowledge available for the modelled system.

The predictive power (via model validation) of kinetic models is often unexplored, usually due to lack of data. The ability of the model to explain experimental data used for parameter estimation cannot guarantee model validation. In this study, we use three independent datasets for validation. The high performance of the dual approach in validation suggests a promising parameter estimation method with respect to model robustness. Our validation results indicate an enhanced predictive power of the calibrated model, which could be used as a tool of in-depth understanding of the biological system. The prediction envelopes provided by the dual approach allow us to better describe the noisy datasets that are typical in biotechnological applications. The robustness of the hybrid GRN-growth kinetic model can be further exploited in model-based optimisation and control strategies in bioremediation and biotechnological applications [54].

Supplementary material

*S1 File.*Details of Approaches I and II and detailed results.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. Almquist, J., Cvijovic, M., Hatzimanikatis, V., Nielsen, J., Jirstrand, M.: Kinetic models in industrial biotechnology–improving cell factory performance. Metabolic engineering **24**, 38–60 (2014)
2. Alon, U.: An introduction to systems biology: design principles of biological circuits. Chapman and Hall/CRC (2006)
3. Ballerstedt, H., Volkers, R., Mars, A., Hallsworth, J., Santos, V., Puchalka, J., Van Duuren, J., Eggink, G., Timmis, K., De Bont, J., Wery, J.: Genomotyping of pseudomonas putida strains using p. putida kt2440-based high-density dna microarrays: Implications for transcriptomics studies. Applied Microbiology and Biotechnology **75**(5), 1133–1142 (2007). DOI 10.1007/s00253-007-0914-z
4. Balsa-Canto, E., Alonso, A., Banga, J.: An iterative identification procedure for dynamic modeling of biochemical networks. BMC Systems Biology **4** (2010). DOI 10.1186/1752-0509-4-11
5. Baranyi, J.: Stochastic modelling of bacterial lag phase. International Journal of Food Microbiology **73**(2-3), 203–206 (2002). DOI 10.1016/S0168-1605(01)00650-X
6. Bertoni, G., Marqués, S., De Lorenzo, V.: Activation of the toluene-responsive regulator xylr causes a transcriptional switch between σ^{54} and σ^{70} promoters at the divergent pr/ps region of the tol plasmid. Molecular Microbiology **27**(3), 651–659 (1998). DOI 10.1046/j.1365-2958.1998.00715.x
7. Bertoni, G., Pérez-Martín, J., De Lorenzo, V.: Genetic evidence of separate repressor and activator activities of the xylr regulator of the tol plasmid, pww0, of pseudomonas putida. Molecular Microbiology **23**(6), 1221–1227 (1997). DOI 10.1046/j.1365-2958.1997.3091673.x
8. Campbell, K., Xia, J., Nielsen, J.: The impact of systems biology on bioprocessing. Trends in Biotechnology **35**(12), 1156–1168 (2017). DOI 10.1016/j.tibtech.2017.08.011
9. Chakrabarti, A., Miskovic, L., Soh, K., Hatzimanikatis, V.: Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. Biotechnology Journal **8**(9), 1043–1057 (2013). DOI 10.1002/biot.201300091
10. Chen, C., Le, H., Goudar, C.: Integration of systems biology in cell line and process development for biopharmaceutical manufacturing. Biochemical Engineering Journal **107**, 11–17 (2016). DOI 10.1016/j.bej.2015.11.013
11. Chen, W., Niepel, M., Sorger, P.: Classic and contemporary approaches to modeling biochemical reactions. Genes and Development **24**(17), 1861–1875 (2010). DOI 10.1101/gad.1945410
12. Cowles, C., Nichols, N., Harwood, C.: Benr, a xy1s homologue, regulates three different pathways of aromatic acid degradation in pseudomonas putida. Journal of Bacteriology **182**(22), 6339–6346 (2000). DOI 10.1128/JB.182.22.6339-6346.2000
13. Cuskey, S., Sprenkle, A.: Benzoate-dependent induction from the op2 operator-promoter region of the tol plasmid pwwo in the absence of known plasmid regulatory genes. Journal of bacteriology **170**(8), 3742–3746 (1988)
14. Dennis, J.E., Gay, D.M., Welsch, R.E.: Algorithm 573: NL2SOL–An Adaptive Nonlinear Least-Squares Algorithm. ACM Transactions on Mathematical Software **7**(3), 369–383 (1981)
15. Devos, D., Garmendia, J., De Lorenzo, V., Valencia, A.: Deciphering the action of aromatic effectors on the prokaryotic enhancer-binding protein xylr: A structural model of its n-terminal domain. Environmental Microbiology **4**(1), 29–41 (2002). DOI 10.1046/j.1462-2920.2002.00265.x
16. Domínguez-Cuevas, P., Marín, P., Busby, S., Ramos, J., Marqués, S.: Roles of effectors in xyls-dependent transcription activation: Intramolecular domain derepression and dna binding. Journal of Bacteriology **190**(9), 3118–3128 (2008). DOI 10.1128/JB.01784-07
17. Egea, J.A., Henriques, D., Cokelaer, T., Villaverde, A.F., Banga, J.R., Saez-Rodriguez, J.: MEIGO: an open-source software suite based on metaheuristics for global optimization in systems biology and bioinformatics. BMC bioinformatics **15**(1), 136 (2013)
18. Egea, J.A., Martí, R., Banga, J.R.: An evolutionary method for complex-process optimization. Computers and Operations Research **37**(2), 315–324 (2010)
19. Ewering, C., Heuser, F., Benölken, J., Brämer, C., Steinbüchel, A.: Metabolic engineering of strains of ralstonia eutropha and pseudomonas putida for biotechnological production of 2-methylcitric acid. Metabolic Engineering **8**(6), 587–602 (2006). DOI 10.1016/j.ymben.2006.05.007
20. Ferrell Jr., J., Tsai, T.C., Yang, Q.: Modeling the cell cycle: Why do certain circuits oscillate? Cell **144**(6), 874–885 (2011). DOI 10.1016/j.cell.2011.03.006
21. Fröhlich, F., Theis, F.J., Rädler, J.O., Hasenauer, J.: Parameter estimation for dynamical systems with discrete events and logical operations. Bioinformatics **33**(7), 1049–1056 (2016)
22. Gábor, A., Villaverde, A.F., Banga, J.R.: Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems. BMC Systems Biology (2017). DOI 10.1186/s12918-017-0428-y
23. Geier, F., Fengos, G., Felizzi, F., Iber, D.: Analyzing and constraining signaling networks: parameter estimation for the user. In: Computational Modeling of Signaling Networks, pp. 23–39. Springer, New Jersey, USA (2012)

24. Gábor, A., Banga, J.: Robust and efficient parameter estimation in dynamic models of biological systems. BMC Systems Biology **9**(1) (2015). DOI 10.1186/s12918-015-0219-2
25. Kiparissides, A., Koutinas, M., Kontoravdi, C., Mantalaris, A., Pistikopoulos, E.: 'closing the loop' in biological systems modeling-from the in silico to the in vitro. Automatica **47**(6), 1147–1155 (2011). DOI 10.1016/j.automatica.2011.01.013
26. Kontoravdi, C., Pistikopoulos, E.N., Mantalaris, A.: Systematic development of predictive mathematical models for animal cell cultures. Computers & Chemical Engineering **34**(8), 1192–1198 (2010)
27. Koutinas, M., Kiparissides, A., Silva-Rocha, R., Lam, M.C., Martins dos Santos, V., de Lorenzo, V., Pistikopoulos, E., Mantalaris, A.: Linking genes to microbial growth kinetics-an integrated biochemical systems engineering approach. Metabolic Engineering **13**(4), 401–413 (2011). DOI 10.1016/j.ymben.2011.02.001
28. Koutinas, M., Lam, M.C., Kiparissides, A., Silva-Rocha, R., Godinho, M., Livingston, A.G., Pistikopoulos, E.N., De Lorenzo, V., Dos Santos, V.A.M., Mantalaris, A.: The regulatory logic of m-xylene biodegradation by pseudomonas putida mt-2 exposed by dynamic modelling of the principal node ps/pr of the tol plasmid. Environmental microbiology **12**(6), 1705–1718 (2010)
29. Kovárová-Kovar, K., Egli, T.: Growth kinetics of suspended microbial cells: From single-substrate- controlled growth to mixed-substrate kinetics. Microbiology and Molecular Biology Reviews **62**(3), 646–666 (1998)
30. Lee, J., Kim, T., Jang, Y.S., Choi, S., Lee, S.: Systems metabolic engineering for chemicals and materials. Trends in Biotechnology **29**(8), 370–378 (2011). DOI 10.1016/j.tibtech.2011.04.001
31. Li, D., Yan, Y., Ping, S., Chen, M., Zhang, W., Li, L., Lin, W., Geng, L., Liu, W., Lu, W., et al.: Genome-wide investigation and functional characterization of the β-ketoadipate pathway in the nitrogen-fixing and root-associated bacterium pseudomonas stutzeri a1501. BMC microbiology **10**(1), 36 (2010)
32. Ljung, L., Chen, T.: Convexity issues in system identification. IEEE International Conference on Control and Automation, ICCA pp. 1–9 (2013). DOI 10.1109/ICCA.2013.6565206
33. Marques, S., Holtel, A., Timmis, K., Ramos, J.: Transcriptional induction kinetics from the promoters of the catabolic pathways of tol plasmid pww0 of pseudomonas putida for metabolism of aromatics. Journal of Bacteriology **176**(9), 2517–2524 (1994). DOI 10.1128/jb.176.9.2517-2524.1994
34. McKellar, R.: A heterogeneous population model for the analysis of bacterial growth kinetics. International Journal of Food Microbiology **36**(2-3), 179–186 (1997). DOI 10.1016/S0168-1605(97)01266-X
35. Moles, C., Mendes, P., Banga, J.: Parameter estimation in biochemical pathways: A comparison of global optimization methods. Genome Research **13**(11), 2467–2474 (2003). DOI 10.1101/gr.1262503
36. Nikel, P.I., Chavarria, M., Danchin, A., de Lorenzo, V.: From dirt to industrial applications: Pseudomonas putida as a synthetic biology chassis for hosting harsh biochemical reactions. Current opinion in chemical biology **34**, 20–29 (2016)
37. Park, J., Lee, S., Kim, T., Kim, H.: Application of systems biology for bioprocess development. Trends in Biotechnology **26**(8), 404–412 (2008). DOI 10.1016/j.tibtech.2008.05.001
38. Pérez-Pantoja, D., Kim, J., Silva-Rocha, R., de Lorenzo, V.: The differential response of the p ben promoter of p seudomonas putida mt-2 to benr and xyls prevents metabolic conflicts in m-xylene biodegradation. Environmental microbiology **17**(1), 64–75 (2015)
39. Pieper, D., Martins Dos Santos, V., Golyshin, P.: Genomic and mechanistic insights into the biodegradation of organic pollutants. Current Opinion in Biotechnology **15**(3), 215–224 (2004). DOI 10.1016/j.copbio.2004.03.008
40. Pitt, J.A., Banga, J.R.: Parameter estimation in models of biological oscillators: an automated regularised estimation approach. BMC Bioinformatics **20**(1) (2019). DOI 10.1186/s12859-019-2630-y. URL `https://doi.org/10.1186/s12859-019-2630-y`
41. Ramos, J., Marqués, S., Timmis, K.: Transcriptional control of the pseudomonas tol plasmid catabolic operons is achieved through an interplay of host factors and plasmid-encoded regulators. Annual Review of Microbiology **51**, 341–373 (1997). DOI 10.1146/annurev.micro.51.1.341
42. Raue, A., Kreutz, C., Theis, F.J., Timmer, J.: Joining forces of bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. Philosophical Transactions Of The Royal Society A **371**(1984) (2013)
43. Raue, A., Schilling, M., Bachmann, J., Matteson, A., Schelke, M., Kaschek, D., Hug, S., Kreutz, C., Harms, B., Theis, F., Klingmüller, U., Timmer, J.: Lessons learned from quantitative dynamical modeling in systems biology. PLoS ONE **8**(9) (2013). DOI 10.1371/journal.pone.0074335
44. Rogers, J., Reardon, K.: Modeling substrate interactions during the biodegradation of mixtures of toluene and phenol by burkholderia species js150. Biotechnology and Bioengineering **70**(4), 428–435 (2000). DOI 10.1002/1097-0290(20001120)70:4⟨428::AID-BIT8⟩3.0.CO;2-4
45. Seber, G.A., Wild, C.J.: Nonlinear regression. hoboken. New Jersey: John Wiley & Sons **62**, 63 (2003)
46. Serban, R., Hindmarsh, A.C.: CVODES: The Sensitivity-Enabled ODE Solver in SUNDIALS. In: Volume 6: 5th International Conference on Multibody Systems, Nonlinear Dynamics, and Control, Parts A, B, and C, pp. 257–269 (2005)

47. Sidoli, F.R., Mantalaris, A., Asprey, S.P.: Toward global parametric estimability of a large-scale kinetic single-cell model for mammalian cell cultures. Industrial & engineering chemistry research **44**(4), 868–878 (2005)
48. Silva-Rocha, R., De Lorenzo, V.: Broadening the signal specificity of prokaryotic promoters by modifying cis-regulatory elements associated with a single transcription factor. Molecular BioSystems **8**(7), 1950–1957 (2012). DOI 10.1039/c2mb25030f
49. Smallbone, K., Mendes, P.: Large-scale metabolic models: From reconstruction to differential equations. Industrial Biotechnology **9**(4), 179–184 (2013). DOI 10.1089/ind.2013.0003
50. Sobol, I.M.: Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. Mathematics and computers in simulation **55**(1-3), 271–280 (2001)
51. Tarantola, A.: Inverse problem theory: methods for data fitting and model parameter estimation. Elsevier (1987)
52. Timmis, K.: Pseudomonas putida: A cosmopolitan opportunist par excellence. Environmental Microbiology **4**(12), 779–781 (2002). DOI 10.1046/j.1462-2920.2002.00365.x
53. Tsipa, A., Koutinas, M., Pistikopoulos, E., Mantalaris, A.: Transcriptional kinetics of the cross-talk between the ortho-cleavage and tol pathways of toluene biodegradation in pseudomonas putida mt-2. Journal of Biotechnology **228**, 112–123 (2016). DOI 10.1016/j.jbiotec.2016.03.053
54. Tsipa, A., Koutinas, M., Usaku, C., Mantalaris, A.: Optimal bioprocess design through a gene regulatory network – growth kinetic hybrid model: Towards replacing monod kinetics. Metabolic Engineering **48**, 129–137 (2018). DOI 10.1016/j.ymben.2018.04.023
55. Tsipa, A., Koutinas, M., Vernardis, S., Mantalaris, A.: The impact of succinate trace on pww0 and ortho-cleavage pathway transcription in pseudomonas putida mt-2 during toluene biodegradation. Bioresource Technology **234**, 397–405 (2017). DOI 10.1016/j.biortech.2017.03.082
56. Villaverde, A., Banga, J.: Reverse engineering and identification in systems biology: Strategies, perspectives and challenges. Journal of the Royal Society Interface **11**(91) (2014). DOI 10.1098/rsif.2013.0505
57. Weiss, R., Basu, S., Hooshangi, S., Kalmbach, A., Karig, D., Mehreja, R., Netravali, I.: Genetic circuit building blocks for cellular computation, communications, and signal processing. Natural Computing **2**(1), 47–84 (2003). DOI 10.1023/A:1023307812034
58. Wiechert, W., Noack, S.: Mechanistic pathway modeling for industrial biotechnology: challenging but worthwhile. Current opinion in biotechnology **22**(5), 604–610 (2011)
59. Ziehn, T., Tomlin, A.S.: Gui–hdmr–a software tool for global sensitivity analysis of complex models. Environmental Modelling & Software **24**(7), 775–785 (2009)

Table 1 The equations forming the GRN model.

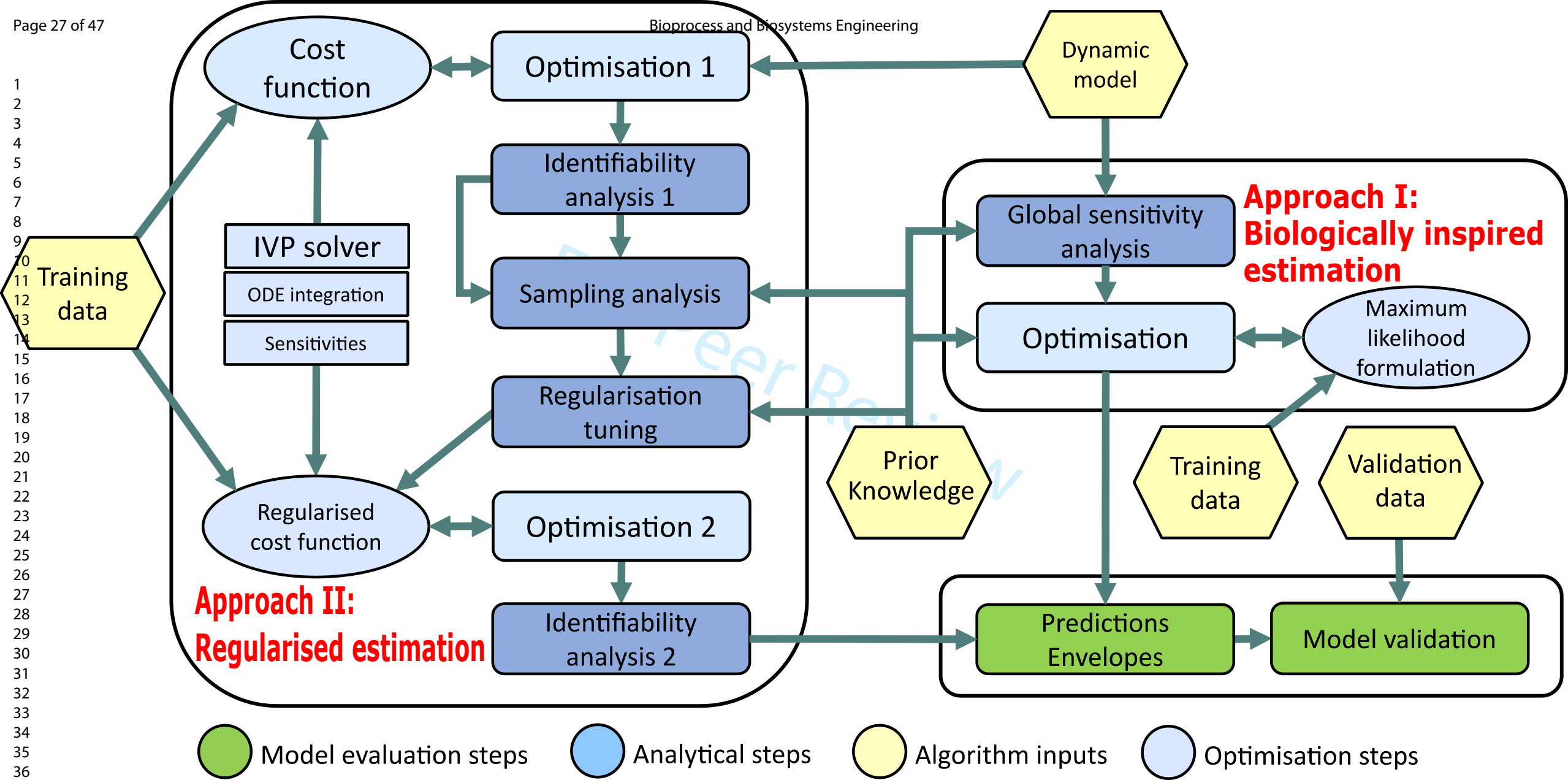
Variables	Equations	
Pr	$\beta_{\mathrm{Pr},tot} = \beta_{\mathrm{Pr}} \left(\frac{K^3_{XylRi}}{K^3_{XylRi} + XylR_i} + \frac{K_{XylRi}}{K_{XylRi} + XylR_a} \right)$	(8)
	$\frac{d\,\mathrm{Pr}}{dt} = \begin{cases} \beta_{\mathrm{Pr},tot} - \alpha_{\mathrm{Pr}} \cdot \mathrm{Pr} & if\, tol \geq 0.3\, mM \\ \beta_{\mathrm{Pr}} \cdot \mathrm{Pr} & otherwise \end{cases}$	(9)
$XylR_i$	$\frac{dXylR_i}{dt} = \beta_{XylRi}\,\mathrm{Pr} - r_{XylR} XylR_i + 3 r_{R,XylR} XylR_a - a_{XylRi} XylR_i$	(10)
	$\frac{dXylR_a}{dt} = \begin{cases} \frac{1}{3} r_{XylR} XylR_i - r_{R,XylR} \cdot XylR_a & if\, t \leq 60\ \mathrm{min} \\ -k_{XylRa} \cdot XylR_a & otherwise \end{cases}$	(11)
r_{XylR}	$r_{XylR} = a \cdot tol_{INI}$	(12)
P_S	$\beta_{Ps,tot} = \beta_{Ps} \frac{XylR_a}{K_{XylRa,Ps} + XylR_a}$	(13)
	$\frac{dPs}{dt} = \beta_{Ps,tot} - \alpha_{Ps} \cdot Ps$	(14)
r_{XylS}	$r_{XylS} = b \cdot tol_{INI}$	(15)
$XylS_i$	$\frac{dXylS_i}{dt} = \beta_{XylSi} \cdot Ps - r_{XylS} \cdot XylS_i + 2 r_{R,XylS} \cdot XylS_a - a_{XylSi} \cdot XylS_i$	(16)
$XylS_a$	$\frac{dXylS_a}{dt} = \frac{1}{2} r_{XylS} \cdot XylS_i - r_{R,XylS} \cdot XylS_a$	(17)
Pu	$\beta_{Pu,tot} = \beta_{Pu} \frac{XylR_a}{K_{XylRa,Pu} + XylR_a}$	(18)
	$\frac{dPu}{dt} = \beta_{Pu} \frac{XylR_a}{K_{XylRa,Pu} + XylR_a} - a_{Pu} \cdot Pu$	(19)
$XylU$	$\frac{dXylU}{dt} = \beta_{XylU} \cdot Pu - a_{XylU} \cdot XylU$	(20)
$BenR_i$	$\frac{dBenR_i}{dt} = \beta_{BenRi} \cdot PbenR - r_{BenR} \cdot BenR_i + 2 r_{R,BenR} \cdot BenR_a - a_{BenRi} \cdot BenR_i$	(21)
r_{Ben_r}	$r_{BenR} = c \cdot tol_{INI}$	(22)
$PbenR$	$\frac{dPbenR}{dt} = \beta_{PbenR} \frac{BenR_a}{K_{BenRa,PbenR} + BenR_a} - a_{PbenR} \cdot PbenR$	(23)
$BenR_a$	$\frac{dBenR_a}{dt} = \begin{cases} \frac{1}{2} r_{BenR} \cdot BenR_i + r_{R,BenR} BenR_a & if\, t \leq 90\ \mathrm{min} \\ -k_{BenRa} \cdot BenR_a & otherwise \end{cases}$	(24)
Pm	$\beta_{Pm,tot} = \beta_{Pm} \cdot \left(\frac{XylS_a}{K_{XylSi,Pm} + XylS_a} + \frac{BenR_a}{K_{BenR,Pm} + BenR_a} \right)$	(25)
	$\frac{dPm}{dt} = \begin{cases} \beta_{Pm,tot} - \alpha_{Pm} \cdot Pm & if\, tol \leq 0.9\, mM\, and\, t < 90\ \mathrm{min} \\ \alpha_1 - \beta_1 \cdot Pm \cdot \frac{R^n_\alpha}{R^n_\alpha + K^n_1} & otherwise \end{cases}$	(26)
$XylM$	$\frac{dXylM}{dt} = \beta_{XylM} \cdot Pm - \alpha_{XylM} \cdot XylM$	(27)
$PbenA$	$\beta_{PbenA,tot} = \beta_{PbenA} \frac{BenR_a}{K_{BenR,PbenA} + BenR_a}$	(28)
	$\frac{dPbenA}{dt} = \beta_{PbenA,tot} - \alpha_{PbenA} \cdot PbenA$	(29)
$BenB$	$\frac{dBenB}{dt} = \beta_{BenB} \cdot PbenA - \alpha_{BenB} \cdot BenB$	(30)

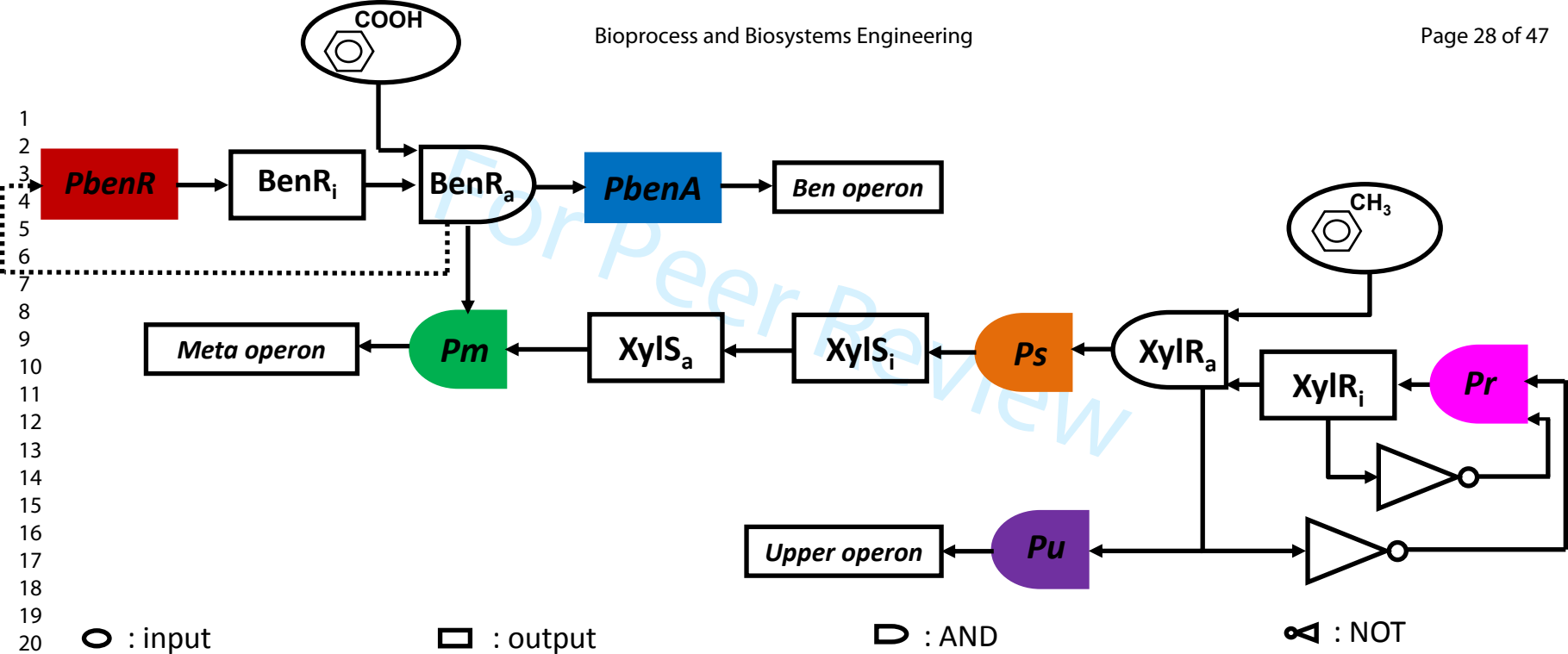
Table 2 Lag phase equation and microbial growth kinetics equations for biomass growth rate, specific growth rate, and toluene utilisation rate.

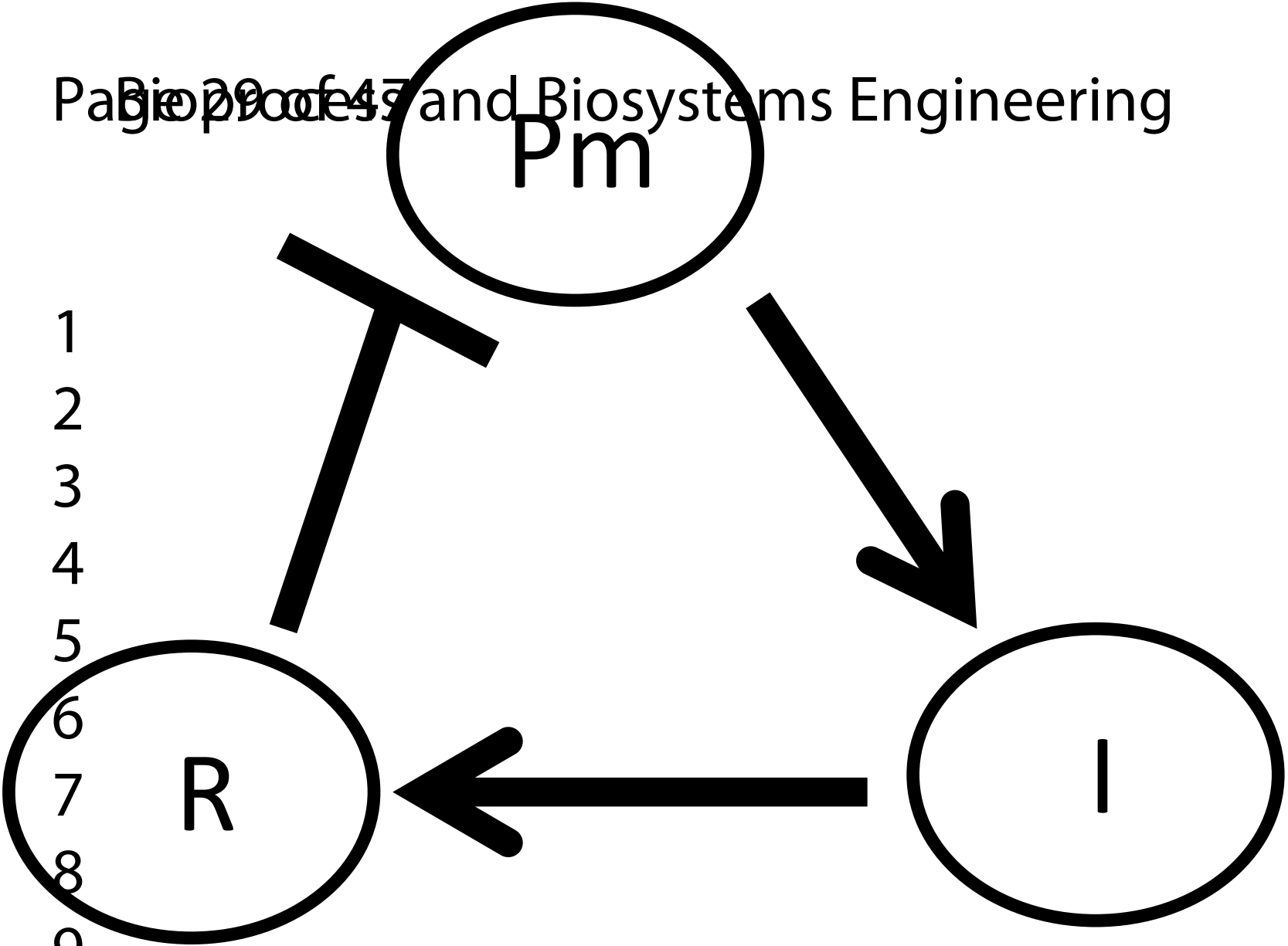
Kinetics	Equations
Toluene	$r_{toluene} = \frac{1}{MW_{toluene}} \frac{\beta_{XylU,toluene} \cdot XylU}{XylU + K_{XylU,toluene}}$ (31)
	$\frac{dtol}{dt} = \begin{cases} 0 & if\ t < lag_{dur}\ and\ tol \leq 0\,mM \\ -r_{toluene} \cdot X & otherwise \end{cases}$ (32)
Specific growth rate	$\mu = \beta_b \frac{XylM}{K_{XylM,b} + XylM} \frac{BenB}{K_{BenB,b} + BenB}$ (33)
Biomass	$\frac{dX}{dt} = (\mu - d) \cdot X$ (34)
	$d = \frac{d_{\max} \cdot tol}{K_d + tol}$ (35)
Oscillatory behaviour of P_m	$\frac{dI_\alpha}{dt} = \frac{1}{2} r_I \cdot I_{in} - r_{R,I} \cdot I_a$ (36)
	$\frac{dI_{in}}{dt} = \alpha_2 \frac{Pm^n}{Pm^n + K_2^n} - r_I \cdot I_{in} + 2 r_{R,I} \cdot I_a - \beta_2 \cdot I_{in}$ (37)
	$\frac{dR_\alpha}{dt} = \frac{1}{2} r_R \cdot R_{in} - r_{R,R} \cdot R_a$ (38)
	$\frac{dR_{in}}{dt} = \alpha_3 \frac{I_\alpha^n}{I_\alpha^n + K_3^n} - r_{R,R} \cdot R_a - \beta_3 \cdot R_{in}$ (39)
Lag phase	$lag_{dur} = l_{TOL} + l_{chrom} \cdot tol_{INI}$ (40)

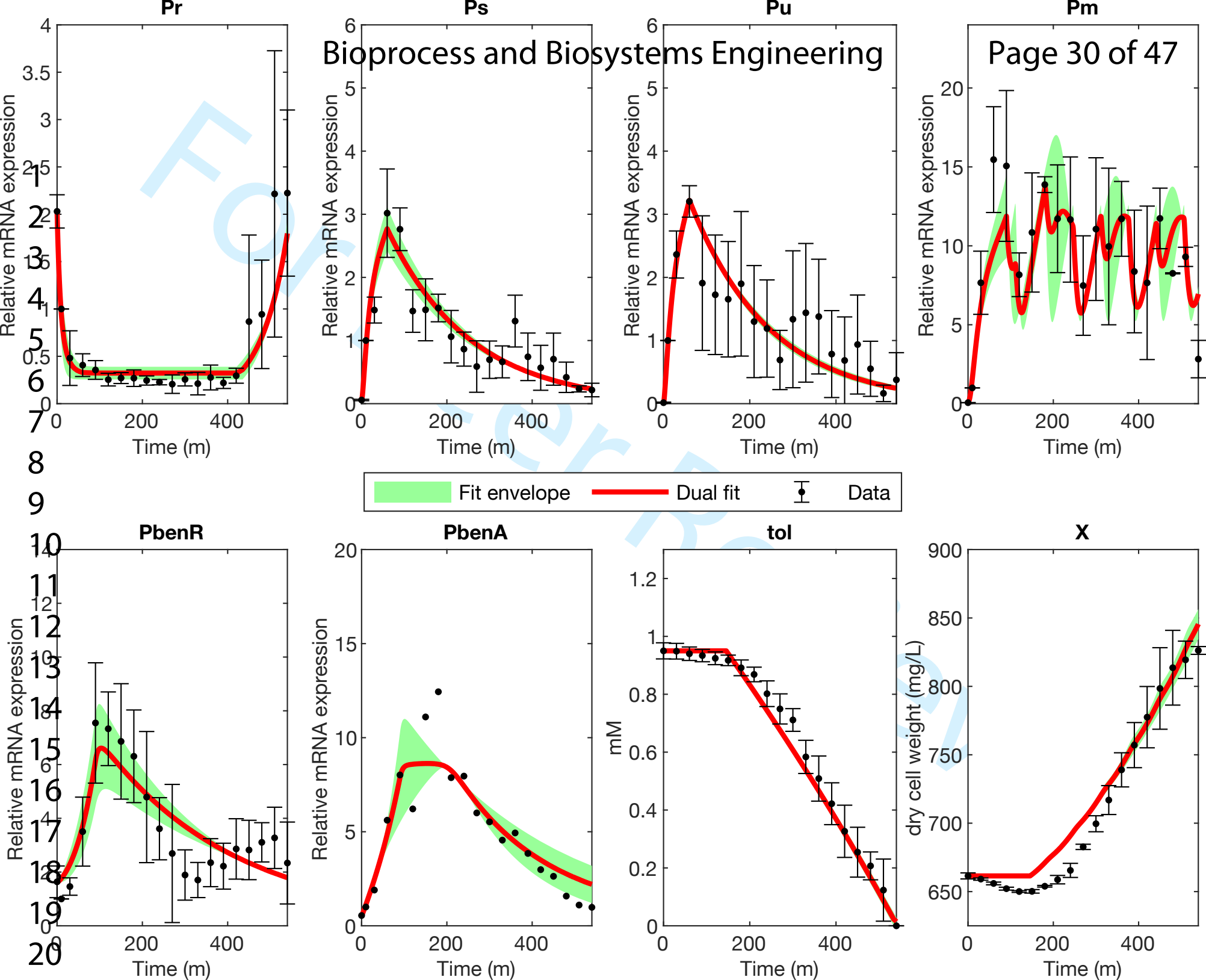
NRMSE table	Calibration	Model validation
Dual approach	0.17055	0.18776
Approach I (biologically inspired estimation)	0.17402	0.25322
Approach II (regularised estimation)	0.19830	0.25227

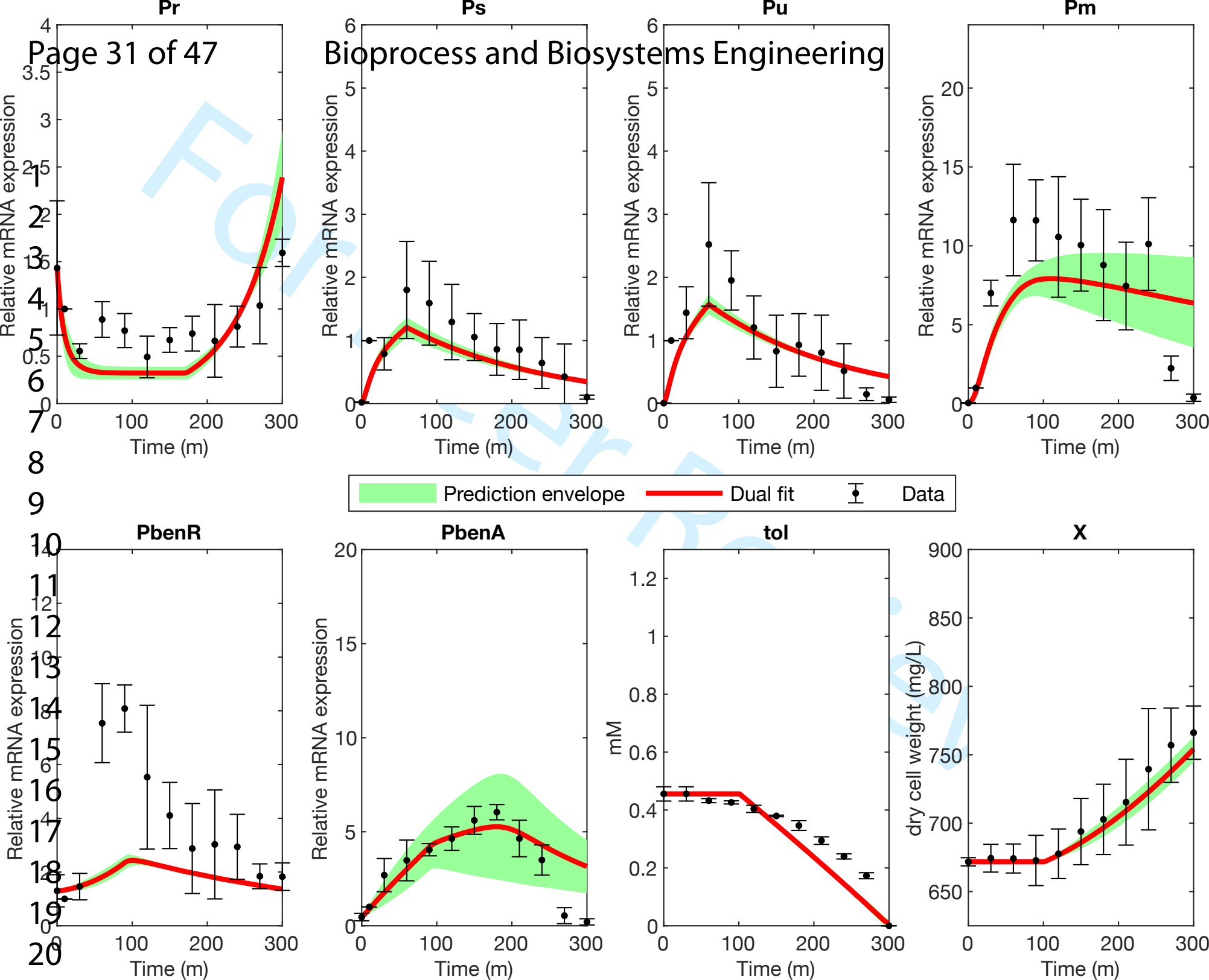
Table 3 Normalised root mean square errors (NRMSEs) of the different approaches for the calibration and validation data sets. Model validation costs are computed considering data from 3 different experiments.

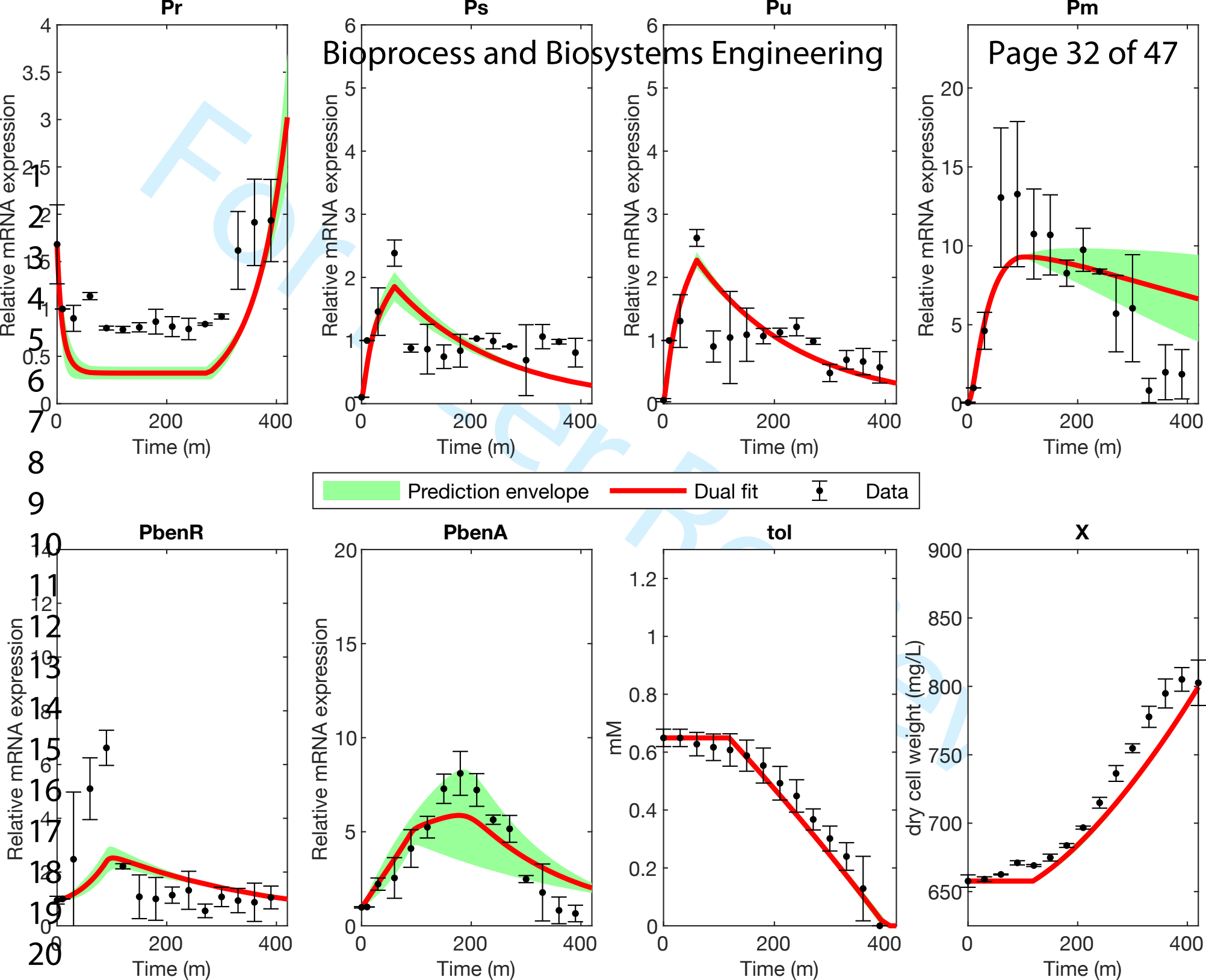

Cost function
Optimisation 1
Dynamic model
Identifiability analysis 1
Training data
IVP solver
ODE integration
Sensitivities
Sampling analysis
Global sensitivity analysis
Approach I: Biologically inspired estimation
Optimisation
Maximum likelihood formulation
Regularisation tuning
Prior Knowledge
Training data
Validation data
Regularised cost function
Optimisation 2
Approach II: Regularised estimation
Identifiability analysis 2
Predictions Envelopes
Model validation
Model evaluation steps
Analytical steps
Algorithm inputs
Optimisation steps

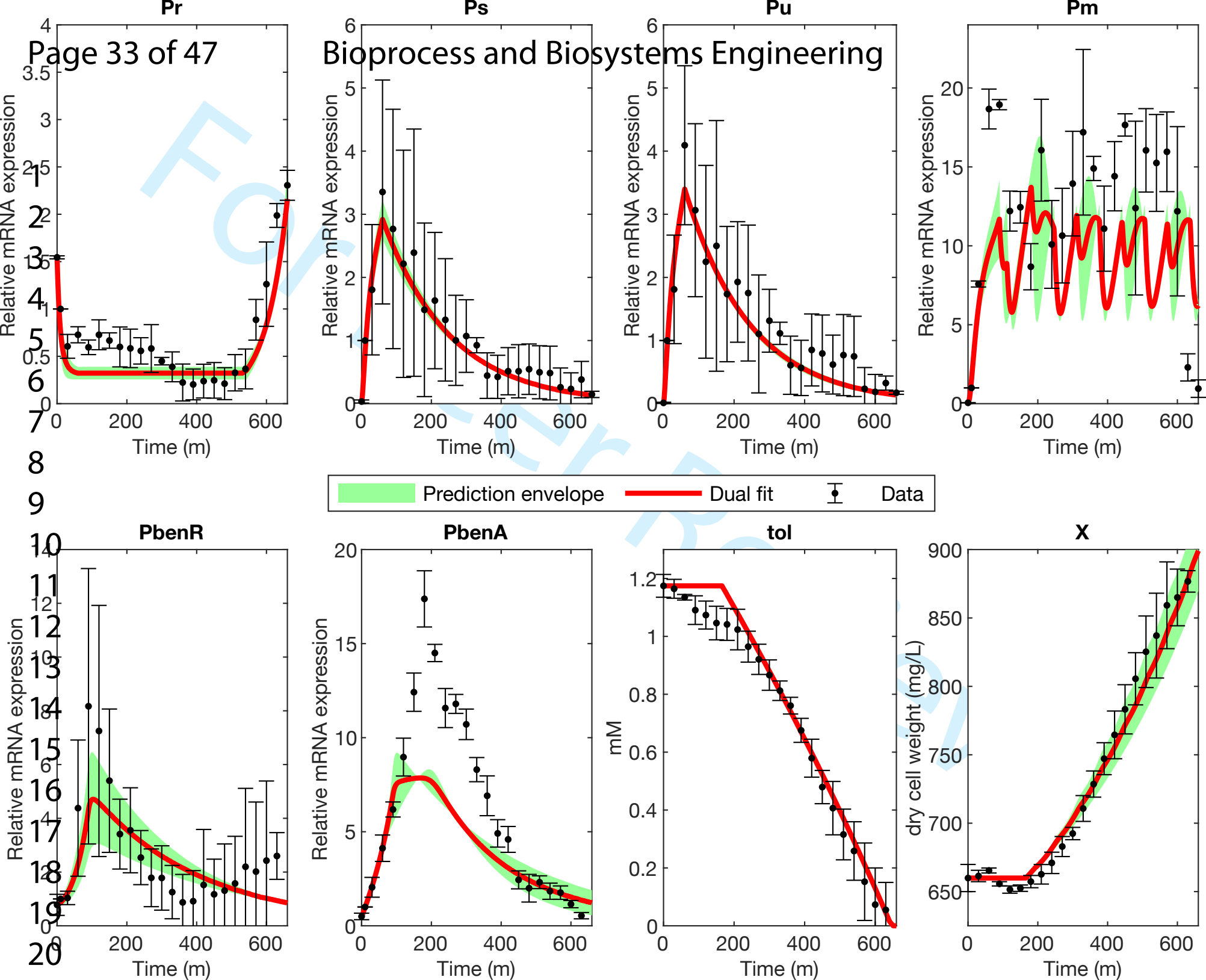



Pm
R
I









S1: Supplementary material for "A dual parameter identification approach for data-driven predictive modelling of hybrid gene regulatory network-growth kinetics in *Pseudomonas putida* mt-2"

Details of Approaches I and II and detailed results.

Argyro Tsipa[1,◕], Jake Alan Pitt[2,3,◕], Julio R. Banga[2,*] and Athanasios Mantalaris[4,*]

[1] Dept. of Civil and Environmental Engineering, University of Cyprus, 75 Kallipoleos Street, 1678 Nicosia, Cyprus
[2] (Bio)Process Engineering Group, Spanish National Research Council, IIM-CSIC, C/Eduardo Cabello 6, 36208 Vigo, Spain.
[3] RWTH-Aachen University Hospital, Joint Research Centre for Computational Biomedicine (JRC-COMBINE), 52074, Aachen, Germany.
[4]Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA.

◕These authors contributed equally to this work.
* Correspondence concerning this article should be addressed to either Prof A. Mantalaris at sakis.mantalaris@gatech.edu, or Prof J. R. Banga at julio@iim.csic.es

S1.1 Approach I: Biologically inspired parameter estimation

S1.1.1 Global sensitivity analysis

S1.1.1.1 GSA for model variables

The 8 observed variables of the model are the promoters of the GRN model: Pr, Ps, Pu, Pm, $PbenR$, $PbenA$, the bioprocess (i.e. biodegradation) variables toluene and biomass. We examined the sensitivity at 50, 70, 100 and 400 minutes. The variables of the GRN were affected by parameters perturbation at 50, 70 and 100 minutes whereas at 400 minutes there is no effect on the GRN variables. This is due to the up-regulatory behaviour of the promoters. Tsipa et al. (2016) [1] noticed that upon toluene entry promoters expression increased and reached maximum levels at 60 or 90 minutes. Following these specific time points, most of the promoters' expression was gradually decreased reaching basal levels. Therefore, the effect of parameters transition on the decline phase does not have any impact on promoters behaviour. The bioprocess variables were affected by parameters perturbation at 400 minutes only. This is because the lag phase lasts for at least 180 minutes for the training data. Therefore, upon toluene degradation and, thus, biomass formation the only point at which parameters transition affect bioprocess design variables is 400 minutes. For the bioprocess variables, the fact that parameters related to GRN are significant underlies the importance of GRN mathematical expression to model microbial growth kinetics.

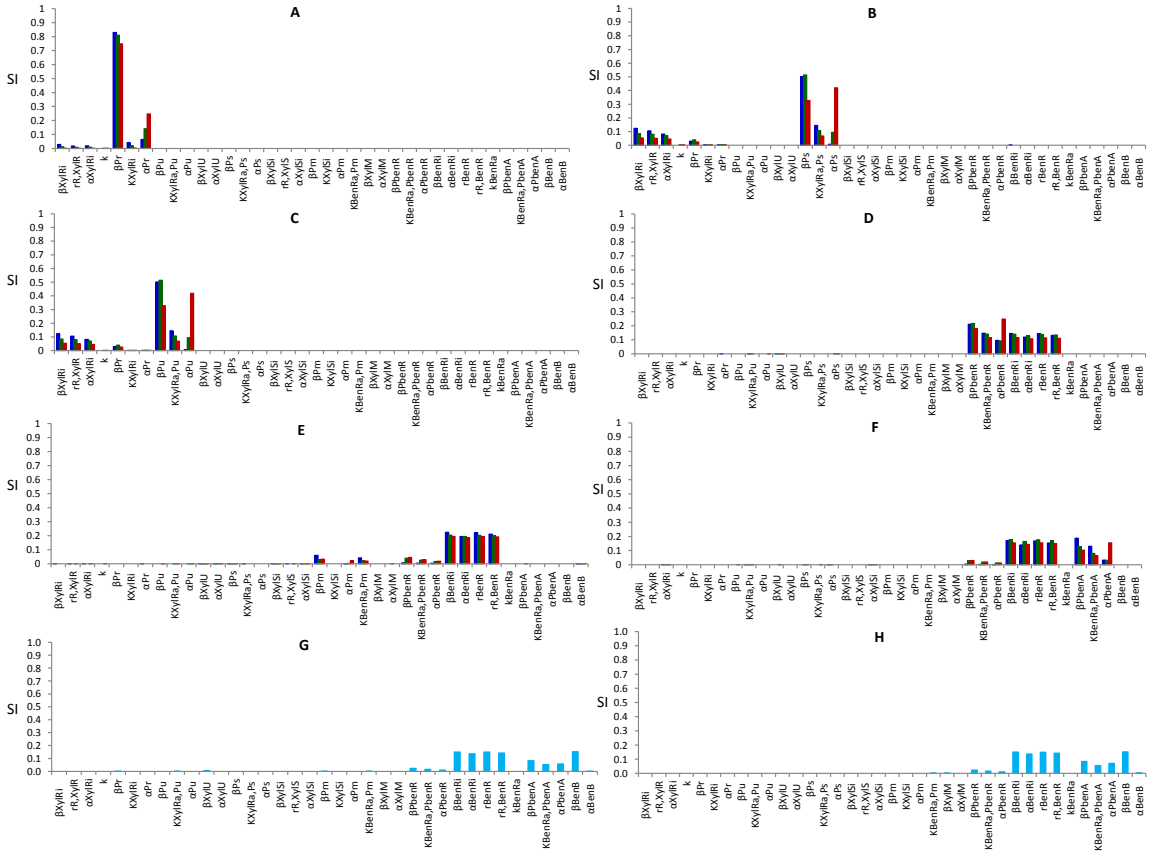


Fig. S1.1 Global sensitivity analysis results of: A)Pr, B)Ps, C)Pu, D)$PbenR$, E)Pm, F)$PbenA$, G)toluene (tol), H) biomass (X) at 50, 70, 100 and 400 minutes

S1.1.1.2 GSA for Pm oscillatory behaviour

The oscillatory behaviour of Pm was consistently observed by Tsipa et al. (2016, 2017) [1, 2] above 0.9 mM toluene concentration threshold and 90 minutes upon toluene entry. However, the transcriptional regulation mechanisms causing this expression pattern is not yet known. Usually, the oscillatory behaviour is caused by participation in negative feedback loop regulation [3]. In an attempt to model this novel behaviour we followed the Ferrell et al. 2011 [4] scenario. In order to check the effect of this model parameter perturbations on the Pm variable, we applied GSA at 110, 130, 140, 160, 170, 190 minutes because at these time points the oscillatory behaviour started to be presented.

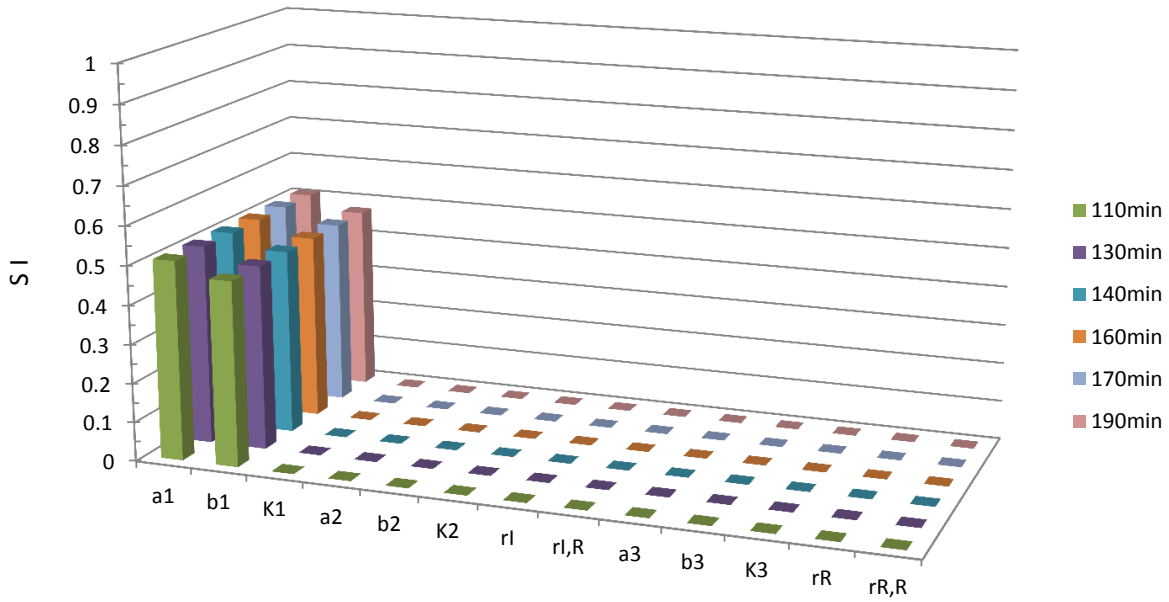


Fig. S1.2 Global sensitivity analysis results of the oscillatory behaviour of *Pm* following the scenario of Ferrell et al (2011)[4]

S1.1.2 Identifiability analysis results

Within approach I we use two different identifiability analyses steps to help reduce the effect of a lack of identifiability on our procedure. The first identifiability analysis is performed after the first optimisation step, using the non-regularised parameter vector to initialise the *VisId* analysis. The results of this analysis are then used to help tune the regularisation procedure as described in S1.7. The second identifiability analysis is the last step of the procedure and is used to update the results of the initial identifiability analysis to account for the fact that we now have achieved the global optima avoiding overfitting. As well as performing the analysis *VisId* [5] also provides a visualisation of the network showing all of the parameters, giving a reason as to why they are not identifiable (Figs S1.3 - S1.4).

Using the `VisId` analysis we find that there are 105 pairwise collinear relationships (Fig. S1.5) and 195 collinear triplets (Fig. S1.6)(S.I), which means that there is a large lack of identifiability within the system. Simply combining parameter values can break these relationships, leading to artefacts. We circumvent this behaviour by instead combining the estimations of our two approaches directly

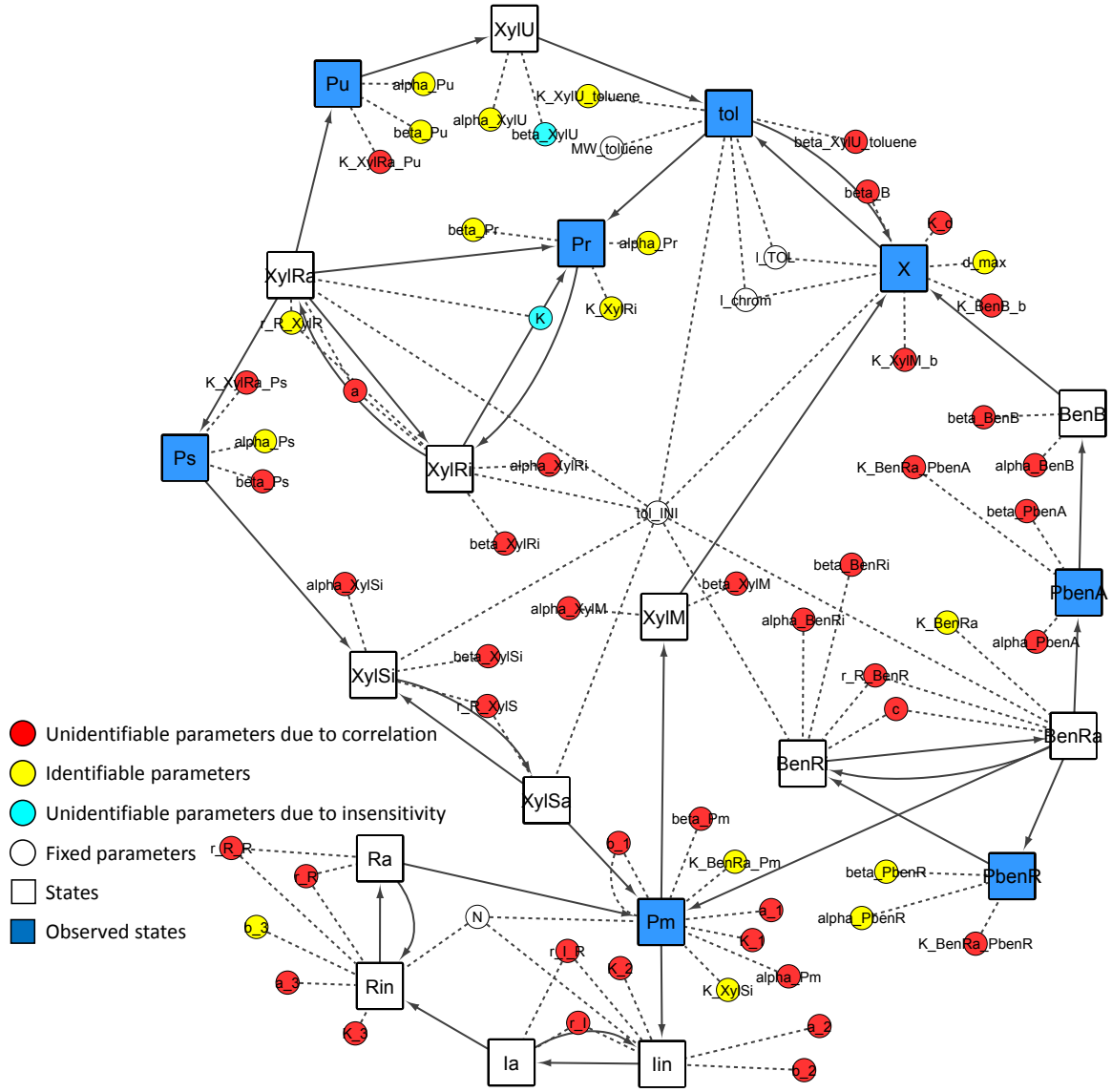


Fig. S1.3 Network representation, showing results of the final identifiability analysis performed. Results from the *VisId* toolbox performed with the non-regularised parameters from the initial estimation.

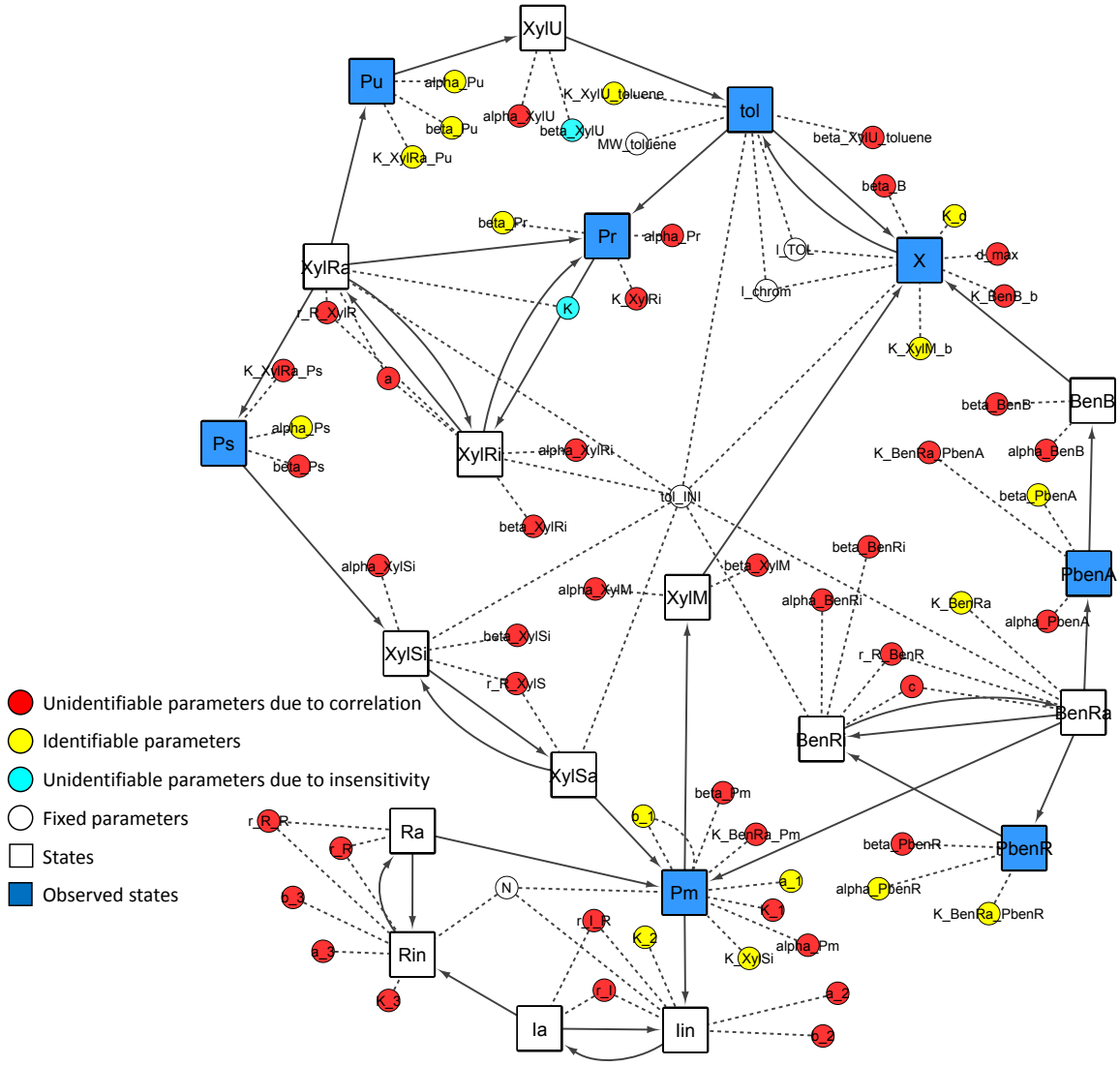


Fig. S1.4 Network representation showing results of the final identifiability analysis performed. Results from the *VisId* toolbox performed with the parameters from the regularised estimation.

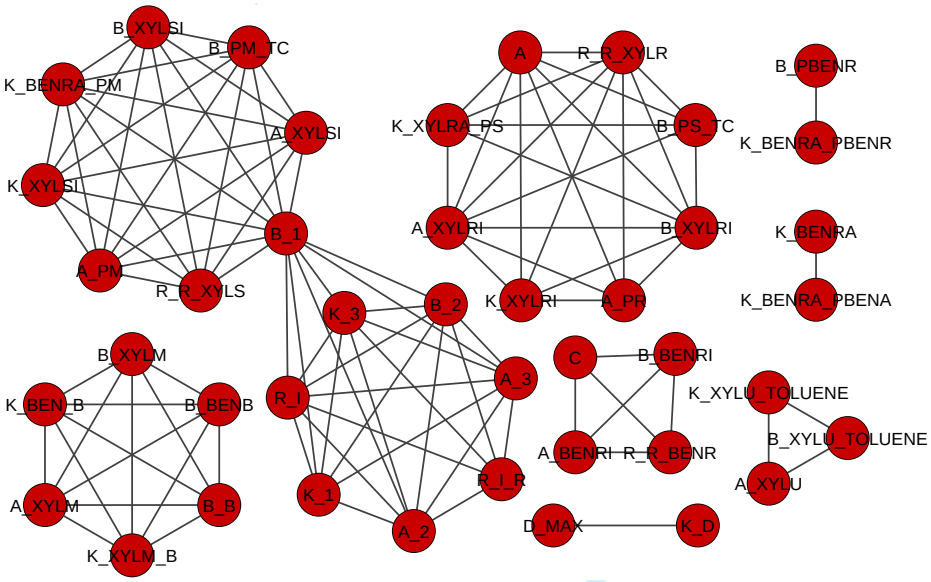


Fig. S1.5 A figure showing the collinear pair relationships. Each edge joins two collinear parameters.

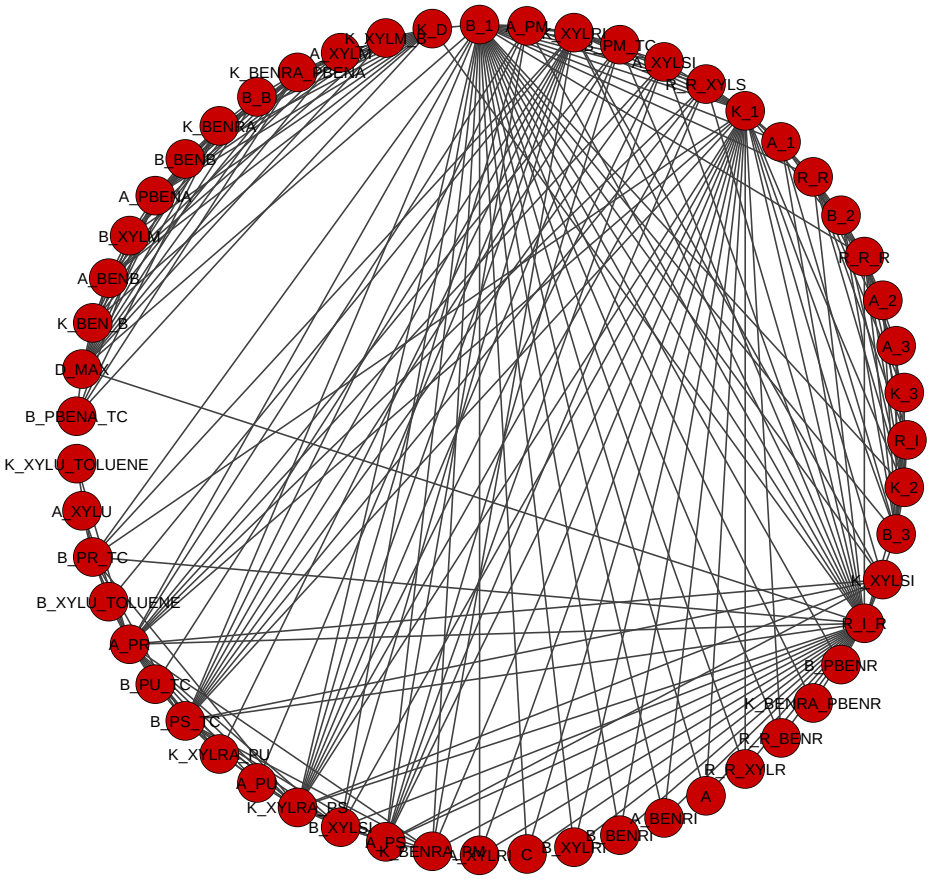


Fig. S1.6 A figure showing the collinear triplet relationships. Each edge joins two collinear parameters.

S1.1.3 Regularisation effect in Approach II

The application of regularisation as described in S1.7 - S1.10 allows for increase predictive power for the regularised estimate over the initial optimisation in approach II. We compare the regularised estimation to the original estimation in approach II both for the calibration (Fig S1.7) and cross-validation sets 1-3 (Figs S1.8 - S1.10 respectively). We also calculate the NRMSE for both estimates (Table S1.1). Without regularisation, we are able to achieve an extremely high-quality calibration (better than the dual approach), but the model has minimal predictive power as can be seen in Figs S1.8 - S1.10 and by the large NRMSE value in Table S1.1. This is a perfect illustration of overfitting. After the application of our regularisation techniques, we are able to sacrifice some of the quality of fit to our calibration data but gain high predictive power.

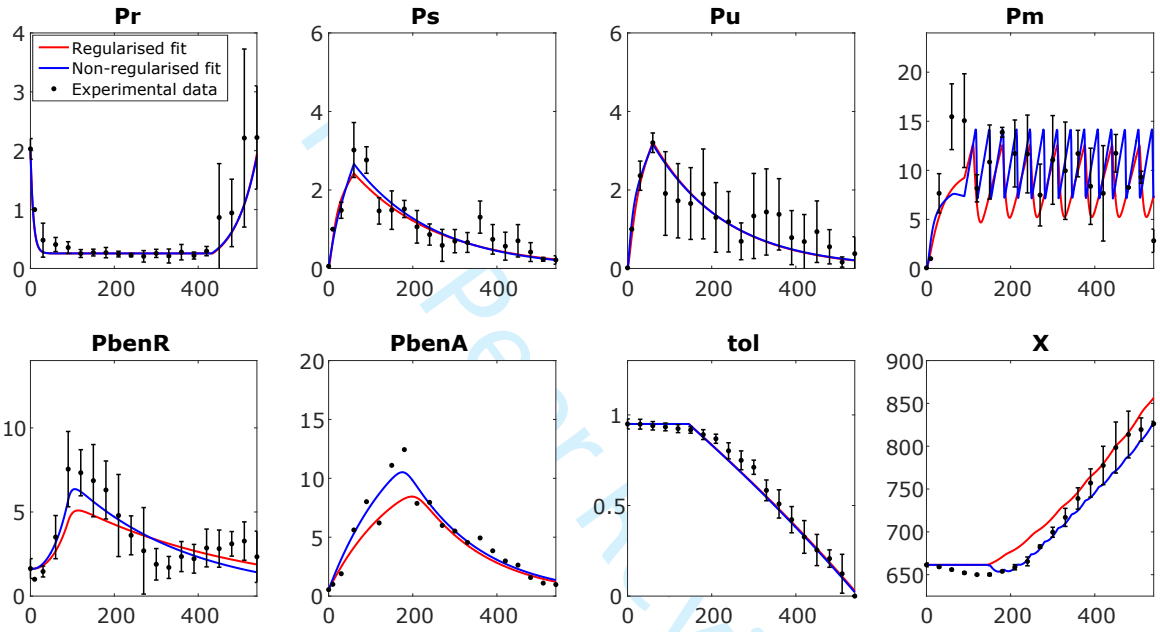


Fig. S1.7 Figure showing the difference between the regularised and non-regularised fits to the calibration data as per the costs described in table S1.1.

NRMSE table	Calibration	Cross-validation
Regularised estimation	0.19830	0.25227
Non-regularised estimation	0.16936	1.45720

Table. S1.1 A table showing the NRMSE for the regularised and non-regularised fits obtained within the regularised estimation process for both the fit and the total validation.

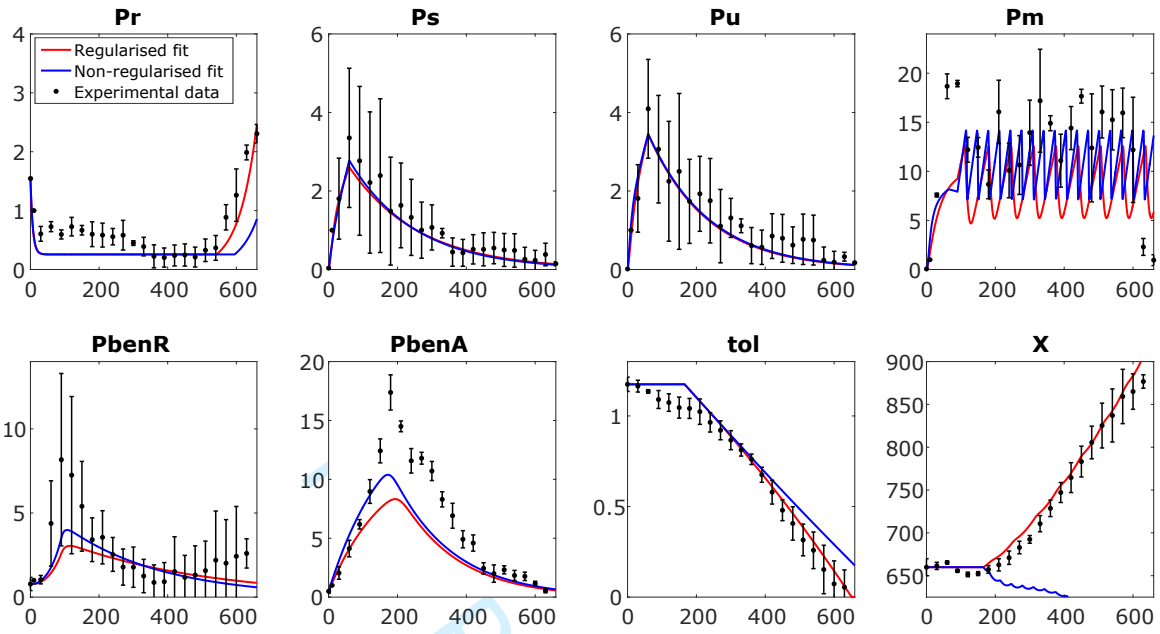


Fig. S1.8 Comparison between the validation of the regularised and non-regularised to validation data set 1 (initial toluene level of 1.2 mM). With the total cost of validation described in table S1.1.

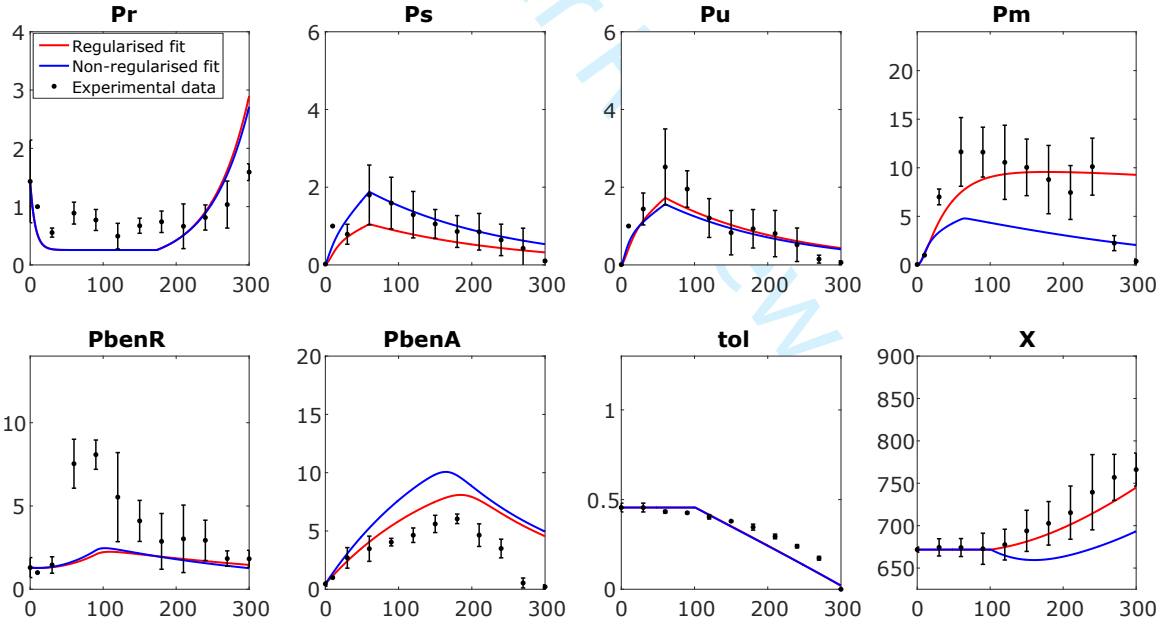


Fig. S1.9 Comparison between the validation of the regularised and non-regularised to validation data set 2 (initial toluene level of 0.4 mM). With the total cost of validation described in table S1.1.

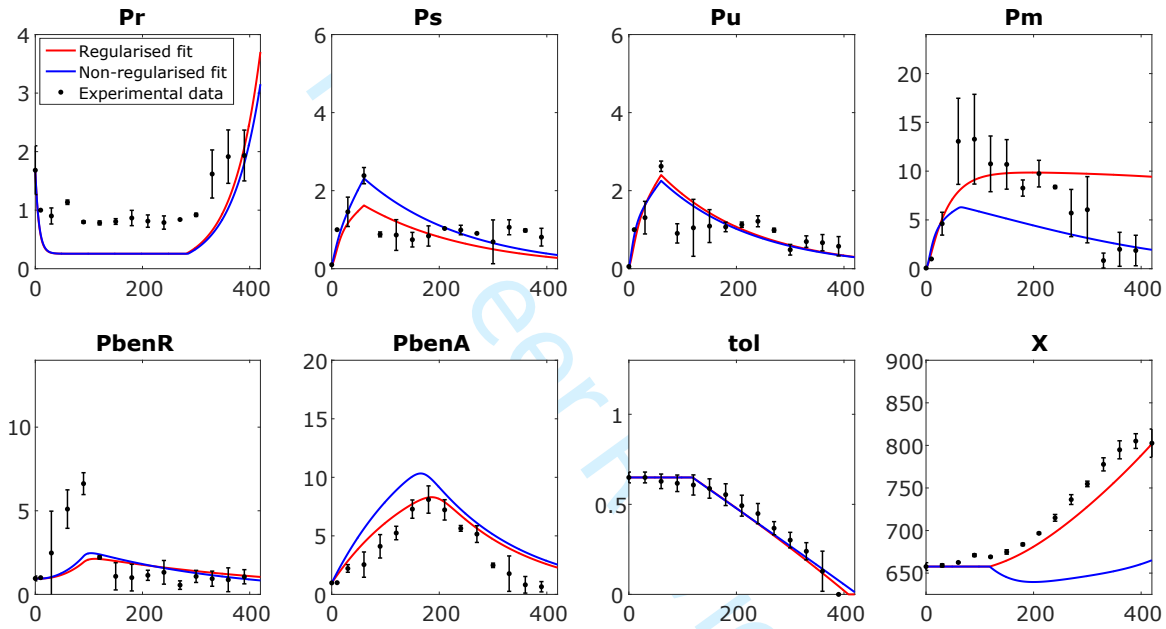


Fig. S1.10 Comparison between the validation of the regularised and non-regularised to validation data set 3 (initial toluene level of 0.6 mM). With the total cost of validation described in table S1.1.

S1.1.4 Parameter vectors from the individual approaches

The two solutions from the parameter estimation approaches can be found in the table below.

Parameter name	Approach I biologically inspired estimation	Approach II regularised estimation
a	7.74271	0.775625222
a_1	0.152491	0.168199183
a_2	1.51061	82.4391255
a_3	1.62035	14.8125508
α_{BenB}	8e-05	6.51597344e-06
α_{BenRi}	1	2.66985584e-05
α_{PbenA}	0.00284994	0.00602623481
α_{PbenR}	0.00361691	0.00240535091
α_{Pm}	0.0549248	0.0288838522
α_{Pr}	0.062208	0.150848182
α_{Ps}	0.00537169	0.00491833681
α_{Pu}	0.00503794	0.00576533383
α_{XylM}	8e-05	0.631321182
α_{XylRi}	1	0.151092649
α_{XylSi}	0.5	0.133277372
α_{XylU}	0.000527257	0.00678918751
b	0.9	4.4556463
b_1	48	0.104611744
b_2	8.47047	13.3886792
b_3	5.05481	0.759877063
β_B	0.00077271	0.0029207548
β_{BenB}	0.555616	0.907228582
β_{BenRi}	4.83041	34.1501619
β_{PbenA}	0.791	0.0744090077
β_{PbenR}	0.742	4.16613541
β_{Pm}	0.721754	0.29309718
β_{Pr}	0.0121162	0.0192946582
β_{Ps}	0.48757	0.574618637
β_{Pu}	0.50365	0.181164739
β_{XylM}	0.397205	14.2635783
β_{XylRi}	9.50432	260.115838
β_{XylSi}	44.5318	31.9428909
β_{XylU}	1.92127	4.34766828
$\beta_{XylU_{toluene}}$	0.000308025	0.000289179356
c	0.16384	0.0090523081
d_{max}	0.00467536	9.65802153e-05
K	17.4367	18295.6745
K_1	0.118888	0.266611598
K_2	5.47328	11.7701024
K_3	13.0198	1.44351198
K_{BenB_b}	0.41501	1534.8

K_{BenRa}	0.203456	0.0843111965
$K_{BenRa_{PbenA}}$	8.63524	0.000102215101
$K_{BenRa_{PbenR}}$	12.3786	87.481524
$K_{BenRa_{Pm}}$	22.6492	762.174155
K_d	12.95	0.261739428
K_{XylM_b}	25.103	195.576429
$K_{XylRa_{Ps}}$	17.2511	49.7971953
$K_{XylRa_{Pu}}$	17.2511	7.59810984
K_{XylRi}	11.0352	1322.41854
K_{XylSi}	33.1017	2.46250635
$K_{XylU_{toluene}}$	3.86312	0.950826018
r_I	7.59798	1.67802653
r_{I_R}	0.109168	2.86783464
r_R	8.2944	0.0907116639
$r_{R_{BenR}}$	0.891	16.6962852
r_{R_R}	0.063118	0.0440634323
$r_{R_{XylR}}$	6.9984	39.5033604
$r_{R_{XylS}}$	0.63133	3.73605563

Table. S1.2 A table showing the parameter values resulting from the solution of optimisations in Approaches I and II.

S1.1.5 NRMSE values of the variables for each approach

Observable-wise NRMSEs Fitting	Approach I biologically inspired estimation	Approach II regularised estimation	Dual approach
Pr	0.15257	0.15413	0.14436
Ps	0.12148	0.10708	0.10317
Pu	0.11977	0.12922	0.12289
Pm	0.15678	0.21603	0.13971
PbenR	0.16684	0.1949	0.1481
PbenA	0.14941	0.12349	0.10965
toluene (tol)	0.052008	0.037484	0.043672
biomass (X)	0.07342	0.083351	0.072549

Table. S1.3 A table showing the NRMSE for each of the methods calculated for each individual observable in the calibration experiment (initial toluene level of 1 mM).

Observable-wise NRMSEs Validation exp 2	Approach I biologically inspired estimation	Approach II regularised estimation	Dual approach
Pr	0.22528	0.4918	0.34712
Ps	0.16596	0.24395	0.20286
Pu	0.18611	0.1469	0.1653
Pm	0.3194	0.32351	0.28207
PbenR	0.35935	0.38773	0.37341
PbenA	0.3233	0.42889	0.22792
toluene (tol)	0.1303	0.10611	0.11795
biomass (X)	0.046155	0 .15622	0.098102

Table. S1.4 A table showing the NRMSE for each of the methods calculated for each individual observable in the first cross-validation experiment (initial toluene level of 0.4 mM).

Observable-wise NRMSEs Validation exp 3	Approach I biologically inspired estimation	Approach II regularised estimation	Dual approach
Pr	0.41632	0.51913	0.4607
Ps	0.18701	0.1864	0.17979
Pu	0.14021	0.15735	0.14791
Pm	0.18378	0.32498	0.24263
PbenR	0.23862	0.27532	0.25547
PbenA	0.28813	0.19488	0.16606
toluene (tol)	0.071077	0.053557	0.061126
biomass (X)	0.11335	0.12287	0.11764

Table. S1.5 A table showing the NRMSE for each of the methods calculated for each individual observable in the second cross-validation experiment (initial toluene level of 0.6 mM).

Observable-wise NRMSEs Validation exp 1	Approach I biologically inspired estimation	Approach II regularised estimation	Dual approach
Pr	0.12129	0.1497	0.12697
Ps	0.055988	0.089403	0.065574
Pu	0.073406	0.083678	0.077684
Pm	0.26107	0.346	0.28939
PbenR	0.15974	0.24372	0.18095
PbenA	0.21412	0.20807	0.20762
toluene (tol)	0.046795	0.053439	0.049501
biomass (X)	0.099545	0.048165	0.047068

Table. S1.6 A table showing the NRMSE for each of the methods calculated for each individual observable in the third cross-validation experiment (initial toluene level of 1.2 mM).

References

[1] A. Tsipa, M. Koutinas, E. N. Pistikopoulos, A. Mantalaris, Transcriptional kinetics of the cross-talk between the ortho-cleavage and tol pathways of toluene biodegradation in pseudomonas putida mt-2, Journal of biotechnology 228 (2016) 112–123.

[2] A. Tsipa, M. Koutinas, S. I. Vernardis, A. Mantalaris, The impact of succinate trace on pww0 and ortho-cleavage pathway transcription in pseudomonas putida mt-2 during toluene biodegradation, Bioresource Technology 234 (2017) 397–405.

[3] U. Alon, An introduction to systems biology: design principles of biological circuits, CRC press, 2006.

[4] J. E. Ferrell, T. Y.-C. Tsai, Q. Yang, Modeling the cell cycle: why do certain circuits oscillate?, Cell 144 (6) (2011) 874–885.

[5] A. Gábor, A. F. Villaverde, J. R. Banga, Parameter identifiability analysis and visualization in large-scale kinetic models of biosystems, BMC systems biology 11 (1) (2017) 54.